FACULDADE DE ECONOMIA DA UNIVERSIDADE
DO PORTO

Dissertation of the Master in Quantitative Methods for
Economics and Management

# Performance Criteria to Validate Simulation Models

*Author:*
Joana da Hora Martins
*Supervisor:*
Professor Pedro Campos

September 28, 2012

**Abstract**

*This thesis explores performance criteria adequate to validate simulation models. An overview on the most widely performance criteria used in literature is firstly provided. The thesis proceeds with the proposal of two new criteria that assess the distortion of the warping path obtained after applying the Dynamic Time Warping algorithm: the Warping Path Distortion (WPD) and the Percentage Warping Path Distortion (PWPD). A case study focused on the demographic evolution of Portuguese firms is presented, whose results are used to perform a comparative analysis on all the criteria revised. This work concludes with a concise outline on the criteria advantages and drawbacks. The criteria WPD and PWPD returned adequate evaluations, yet further applicability of these measures to benchmark data sets is necessary to provide a proper conclusion on its quality.*

**Keywords:** performance criteria, validation, simulation models, Warping Path Distortion, Percentage Warping Path Distortion.

*"We have described the principle of induction as the means whereby science decides upon truth. To be more exact, we should say that it serves to decide upon probability. For it is not given to science to reach either truth or falsity . . . but scientific statements can only attain continuous degrees of probability whose unattainable upper and lower limits are truth and falsity."*

Hans Reichenbach (in Erkenntnis 1, 1930, pp. 186)

# Acknowledgements

I am grateful to my supervisor, Professor Pedro Campos, for the knowledge, patience and motivation he granted to me during the development of this thesis.

I want to congratulate all Professors of this Master course, with whom I had the pleasure of improving my knowledge in different subjects.

I leave a word of gratitude to the institution where I work, INESC TEC, for the flexibility and support always demonstrated.

Finally, I want to thank to my dear family and friends, for their love and affection.

# Contents

# List of Figures

# List of Tables

# Nomenclature

$\rho(n)$ Cross correlation

$\tau$ Kendall's Tau

$AIC$ Akaike Information Criterion

$AR$ Autoregressive

$ARMA$ Autoregressive Moving Average

$BIC$ Bayesian Information Criterion

$C_{NISE}$ Normalized Integral Square combined error

$C_R$ Russel's combined error

$C_{S\&G}$ Sprague and Gear combined error

$CR$ Critical Region

$d_r$ Minkowski distance

$D_{KL}$ Kullback-Leibler Divergence

$d_{r=1}$ Manhattan distance

$d_{r=2}$ Euclidean distance

$d_{STS}$ Short Time Series distance

$DIC$ Deviance Information Criterion

$DTW$ Dynamic Time Warping

$EIC$ Extended Information Criterion

$FIC$ Focused Information Criterion

$GIC$ Generalized Information Criterion

$H_{r2}$ Quadratic Entropy described by Renyi

$H_{r\alpha}$ General Entropy described by Renyi

$H_S$ Entropy described by Shannon

$I$ Mutual Information

$IC$ Information Criteria

$Id$ Normalized Information Distance

$IT$ Information Theory

$M_{NISE}$ Normalized Integral Square magnitude error

$M_R$ Russel's magnitude error

$M_{S\&G}$ Sprague and Gear magnitude error

$MA$ Moving Average

$MAE$ Mean Absolute Error

$MAICE$ Minimum Information Theoretic Criterion Estimate

$MAPE$ Mean Absolute Percentage Error

$MASE$ Mean Absolute Scaled Error

$ME$ Mean Error

$MPE$ Mean Percentage Error

$MSE$ Mean Square Error

$NIC$ Network Information Criterion

$NMI$ Normalized Mutual Information

$NUTS$ Nomenclature of Territorial Units for Statistics

$P_{NISE}$ Normalized Integral Square phase error

$P_R$ Russel's phase error

$P_{S\&G}$ Sprague and Gear phase error

$PE$ Percentage Error

$PoFi$ Simulation Model to study Portuguese Firms

$PWPD$ Percentage Warping Path Distortion

$r$ Coefficient of correlation

$R^2$ Coefficient of determination

$r_S$ Spearman's rank correlation coefficient

$RMSE$ Relative Mean Absolute Error

$RMSE$ Root Mean Square Error

$S_{NISE}$ Normalized Integral Square shape error

$SIC$ Schwarz Information Criterion

$TIC$ Takeuchi's Information Criterion

$TS$ Test Statistic

$U_1$ Theil's Measure of Forecast Accuracy

$U_2$ Theil's Measure of Forecast Quality

$WPD$ Warping Path Distortion

# Chapter 1

# Introduction

Simulation models developed by applying computational tools are increasingly used to study various problems. The development of simulation models improves the knowledge of the systems under study and supports the decision making process [42, 35, 41]. The developers of simulation models intend to provide information as accurate as possible. The users of these models (decision makers using the resulting information and the individuals affected by the decisions taken) have interest on the correctness of the information attained. Therefore, the correctness level of the results obtained with a simulation model is a basilar issue to be addressed.

The development of a computational simulation model encompasses several phases [41, 35]. The analysis of the system to study from reality provides the necessary information to further specify a conceptual model. The conceptual model is a set of relationships between features, mathematically defined, believed to best traduce the system under analysis. Once the conceptual model is specified, the programming phase is conducted, from where a model code is obtained. The next step is the calibration, which adjusts the parameters of the model to better fit the data from reality. The last phase is the validation, which returns the assessment of the goodness of fit of the model through the calculation of performance criteria using data from reality.

The literature on simulation models includes distinct meanings for the term *validation*, and the concept of validation is normally set alongside with other concepts such as quality [48], verification, confirmation and calibration [41, 35]. There are also critics on the use of the validation term [22] (arguing that a model is not designed to predict but to reproduce the observed historic records, ergo these terms could be replaced by other terms that would clearly indicate that a model is designed to replicate a historic data set).

The main disagreement on the meaning of this term rely on philosophic questions

about whether scientific research should be based on inductive or deductive methods. A perspective that criticizes the inductive method is provided by Karl Popper in his book [37]: "the principle of induction must be a synthetic statement; that is, a statement whose negation is not self-contradictory but logically possible. So the question arises why such a principle should be accepted" ([37], pp. 5). A perspective that defends the inductive method is made by Kuhn: "Few philosophers of science still seek absolute criteria for the verification of scientific theories. Noting that no theory can ever be exposed to all possible relevant tests, they ask not whether a theory has been verified but rather about its probability in the light of the evidence that actually exists. And to answer that question one important school is driven to compare the ability of different theories to explain the evidence at hand" ([24], pp. 145).

The term validation is used in this work with the following meaning: a model is validated when the results obtained return a satisfactory range of accuracy for a specified performance criterion, considering the respective domain of applicability. Noting that a simulation model is constructed always under specific conditions, the knowledge extracted from its results should always be interpreted with reference to the assumptions and hypothesis formulated. The validation of a model is in fact the validation of a set of hypothesis, defined with mathematical formulas and governing relationships over the features considered.

How should a validation process be conducted? How can one conclude on the validity of a simulation model? What are the basis for credibility of a given simulation model? These were the main questions that motivated this work.

The validation process is quite complex, as it involves many delicate aspects which are not deepened in this work. This thesis is focused on the quantitative performance criteria adequate to be used under a validation process.

A performance criteria can be a measure, a metric, a statistical test or an empirical procedure. The performance criteria revised in this work are adequate to assess paired data samples. They return a quantitative assessment on the goodness of fit between estimated and observed data. The criteria revised are adequate to compare more than two data sets, provided there exist comparable homologous elements within the data sets under analysis. Depending on the modeling technique applied, the calibration phase may apply similar performance criteria as the validation phase.

The development of simulation models is limited by the knowledge about reality, by the available data on the features to be simulated, and by the computational effort that can be used. The challenge is then, with the necessary awareness about the limitations that one can found, to develop simulation models that best accomplish the task for which they

were defined.

A model can be accepted as valid for specific conditions only. To conclude that a model is valid, the scientific standard procedures for *validation* and *calibration* should be followed, using data from *reality* to confront the results obtained [41]. The validation outcome is further used to *confirm* or reject a *conceptual model*.

## 1.1 Main goals pursued

The main goals pursued in this work are described as follows.

- To provide a comprehensive bibliographic review on performance criteria adequate to assess the simulation results in comparison with the homologous elements observed from reality.
- Investigate new performance criteria adequate to assess the goodness of fitness of paired data sets.
- Test and compare the performance criteria reviewed using data from a practical example, pointing out the advantages and drawbacks of each criterion.

## 1.2 Main contributions

The main contributions provided by this work, are summarily described next.

- A comprehensive bibliographic survey on performance criteria adequate to validate simulation models is provided.
- Two new performance criteria are proposed: the *Warping Path Distortion* - $WPD$ and the *Percentage Warping Path Distortion* - $PWPD$. These criteria return a measure of pattern similarity between two data sets.
- The performance criteria reviewed are calculated and compared for the results obtained with the simulation model *PoFi*, which was developed by the author using cellular automata thecniques to study demographic aspects of **Po**rtuguese **Fi**rms.

## 1.3 Structure of this thesis

This thesis is structured as follows.

Chapter 2 includes a summary overview on (i) relevant terminology to the research of simulation models, and (ii) the classic methodology to coordinate the processes validation and calibration, alongside with the data split techniques most widely used.

Chapter 3 provides a bibliographic survey on the performance criteria most widely used to validate simulation models. It is organized in seven sections: (i) measures to assess the divergence between two data samples with base on error, (ii) measures based on information theory, (iii) information criteria, (iv) parametric tests and measures, (v) non-parametric tests that may be applicable to paired samples, (vi) distance measures between two vectors, (vii) combined measures resulting from the assessment of magnitude, phase and shape characteristics.

Chapter 4 includes the proposal of two new performance criteria: (i) the Warping Path Distortion- $WPD$ and (ii) the Percentage Warping Path Distortion- $PWPD$.

Chapter 5 starts with a summary description of the *PoFi* model. It proceeds with the analysis of the performance criteria reviewed in chapters 3 and 4 applied to five experiments. These experiments refer to the results obtained with the *PoFi* concerning: the number of firms in *Norte*, *Centro*, *Lisboa*, *Alentejo* and *Algarve*.

The conclusions of this work are presented in Chapter 6.

# Chapter 2

# Terminology and Data aspects

This chapter includes two main sections. The first section provides a description of relevant terminology for the construction of computational simulation models. The second section includes a summary overview on the classic methodologies to coordinate the processes of validation and calibration, alongside with the classic techniques for data split into calibration and validation data sets. Whenever relevant, the support of the example *demographic evolution of Portuguese firms* is used.

## 2.1 Terminology

The work developed by Refsgaard [41] reviews many perspectives on main concepts concerning computational simulation models, and proposes an unifying terminology. The terminology proposed in [41] is adopted in this thesis, and is summarily described. A causal-relationship scheme including the relevant terms is presented in Figure 2.1.

- **Reality**

  The real system to be studied and modeled.

  For example, a set of real Portuguese firms could be a reality to study. From the real firms, it is possible to collect historic data concerning relevant features to study further, such as the number of living firms in each geographic zone over time.

- **Conceptual model**

  A conceptual model includes (i) a mathematical formulation (equations) and (ii) a description on the most relevant features to be simulated. It aims to describe the reality in terms of verbal descriptions, equations, governing relationships or

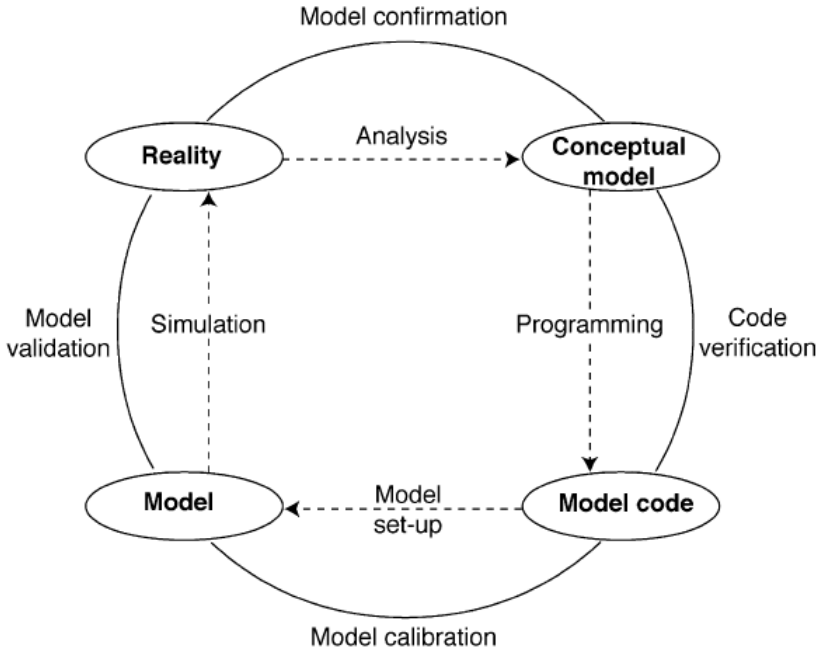


Figure 2.1: Terminology for the construction of computational simulation models (from [41]).

natural laws, using the most accurate perception of the key features to be modeled (perceptual model) and the corresponding simplifications and numerical accuracy limits that are assumed acceptable. A conceptual model constitutes the scientific hypothesis or theory assumed for the model under study.

For example, the hypothesis *the birth rate of firms in a specific moment is exponentially dependent on the number of living firms for that same moment* could be formulated to be further tested. This hypothesis can only be precisely specified provided all variables used (birth rate of firms, number of living firms and time moments) are described unambiguously.

- **Model code**

  A computer program with the implementation of the generic mathematical formulation hypothesized in the conceptual model. The generality of a model code means that it can be used to create distinct models for different case studies, using the same elemental equations and allowing distinct input variables and parameter values.

- **Model**

  The model is established for a particular case study. It is constructed from the model

code, using input data to find parameter values.

- **Code verification**

  The code verification is the task which allows the substantiation that a model code is a suitable representation of a conceptual model, considering specified limits or ranges of application and corresponding ranges of accuracy. This is normally done with a methodological debug of model code, in order to ensure that it performs exactly the desired tasks.

  In the work developed by Gilbert and Troitzsch [16], focused on simulation for social systems, the model verification if defined as "the process of checking that a program does what it was planned to do" ([16], pp.21). Moreover, the following insightful recommendations concerning the verification phase are provided by [16]: (i) the debug should be made carefully and preferably using a set of test cases, such as simulating extremes situations, in order to easily check weather the results of the model are coincident with the expectable results; (ii) it is advised the rerun of the model each time a major change is made, in order to easily check possible errors within the change made; (iii) each rerun of the model should automatically include the whole list of test cases, record the respective results and if possible, highlight major differences over different runs; (iv) finally, a record of the results and the code of each run should be stored within a "version control system".

- **Model confirmation**

  The assessment on the conceptual model adequacy to provide an acceptable level of agreement for the domain of intended application. When this assessment is considered acceptable, the theories / hypotheses included in the conceptual model are scientifically confirmed. Otherwise, the theories / hypotheses are rejected.

- **Model calibration**

  The process of adjustment of parameter values included in the model in order to approximate the model results to reality, considering a specified range of accuracy in the performance criteria used. The term training is frequently applied with similar meaning (e.g. [32]).

  Some modeling approaches differentiate two tasks under the calibration process: the training and the test (noting that the term test can be used both as a part of the calibration process and as the validation process).

- **Model validation**

  A model is validated when, under its domain of applicability, the model outputs return a satisfactory range of accuracy for a specified performance criterion, that should be consistent with the intended application of the model. The term testing is used with the same meaning as validation in some modeling approaches (e.g. in [32]).

- **Model set-up**

  Establishment of a model for a specific case study with base on the model code. This establishment is made with the definition of (i) boundary, (ii) initial conditions and (iii) parameter assessment from field and laboratory data.

- **Simulation**

  Simulation is the use of a validated model to gain insight about reality. Simulation can be used to predict the evolution of some features or to study how reality is expected to evolve over the change of specific features included into the model. It is important to consider the uncertainty underlying the model when defining the conclusions of the studied reality.

- **Analysis**

  The assessment of the model quality considering (i) the reality and (ii) the scientific description of reality (i.e. the conceptual model with its theories and equations). The analysis is performed with analytic tools, and can lead to conclude on a good or bad result for the model confirmation.

  As explained by [16], when simulation models are constructed to reproduce stochastic processes (when the reality to be simulated is at least, partly based on random factors), it is appropriate to perform a sensitivity analysis. A sensitive analysis of a simulation models aims (i) "to answer questions about the extent to which the behavior of the simulation is sensitive to the assumptions which have been made" ([16], pp.23), and (ii) to investigate the robustness of the model. It is conducted by running the model under different values for the initial conditions and parameters (for example by performing small sequential changes or by randomizing these values), with the analysis of the corresponding outcomes. This approaches allows the study of the behavior of the model under different conditions.

- **Programming**

The creation of the model code from the conceptual model, using computational tools.

Another important concept described in [48] refers to the quality of a simulation model, where three main quality goals are identified: (i) *performance*: the ability of a simulation model to execute its roles, with efficiency and reliability; (ii) *safety*: the adequacy of the simulation model to prevent accidents when applied to perform some process in the real world (e.g. controlling a machine); and (iii) *security*: the adequacy of a simulation model to comply with laws, norms and standards (this goal is connected with the topics of confidentiality, integrity or authenticity).

A short summing on the main concepts revised in this section is now presented. A simulation model is designed to validate or reject specific theories and hypothesis considered in the conceptual model, always under the specific conditions and assumptions made. Therefore, a simulation model can be accepted as valid for specific conditions only, which means that under the conditions simulated, the model reproduces well the task for which it was designed. To prove that a model reproduces well its task (i.e. to *confirm* a model) it is important to follow the standard procedures [41], implying the use of data from *reality* to confront the results obtained, and use the *calibration* and *validation* outcomes as valuable information to *confirm* (i.e. the *model* corroborates the hypothesis defined in the *conceptual model*) or reject a *conceptual model*.

## 2.2 Calibration and Validation methodology

A list on different techniques to validate simulation models is provided by [42]. This list includes the following validation techniques: (i) Animation, (ii) Comparison to Other Models, (iii) Degenerate Tests, (iv) Event Validity, (v) Extreme Condition Tests, (vi) Face Validity, (vii) Historical Data Validation, (viii) Historical Methods, (ix) Internal Validity, (x) Multistage Validation, (xi) Operational Graphics, (xii) Parameter Variability - Sensitivity Analysis, (xiii) Predictive Validation and (xiv) Traces and (xv) Turing Tests. Also, an empirical validation approach is suggested in [14]. Although the extensive range of validation approaches available in literature, this thesis is focused on the *Historical Data Validation* one. Accordingly, the performance criteria reviewed in further chapters are suitable to assess quantitatively paired data sets.

The Historical Data Validation is applied when there are historical records or collected data from *reality*. These data is the reference point that the model should be able to

reproduce, and is normally separated in (i) data used to build the model (i.e. to use in model *calibration*) and (ii) data used to test whether the model behaves as the system does (i.e. to use in the model *validation* process). According to [41], the measurement of the level of acceptable agreement between the model and the reality is made with *performance criteria*, both for model calibration and model validation. It is possible to use the same performance criteria in calibration and in validation.

The validation process is deeply connected with calibration. The methodology used to coordinate these two processes is extremely important (the selection of an appropriate performance criterion would be worthless if not properly applied).

Concerning the data to be applied within the processes of validation and calibration, three main attributes are detailed by [32]: (i) the calibration data may provide the model with *direct* or *indirect* information concerning specific aspects the model is intended to learn. Normally, a model learns its tasks easily when provided with *direct* information, and the *credit assignment* problem may arise when the model is provided with *indirect* information (see [32], pp. 5); (ii) The sequence of the training examples provided to the model within the calibration data set is an important aspect to consider (for example the data can be randomly ordered); (iii) the third attribute is the distribution similarity assessment between calibration and validation data sets.

The next sections summarily describe classic methodologies used (i) to split the data to be included within the validation and calibration data sets, and (ii) to coordinate the validation and calibration processes in order to achieve meaningful results. Accordingly, five methods are revised. The first method is the cross validation test, which allows validation on a sample used for model calibration. The second method is the classic split sample test, which defines that the calibration data set is independent from validation data set. The other three methods are variants of the split sample test, all of them prosed by Klemeš [21] to perform split sample test within situations with insufficient data available and to test the model behavior for changes in features.

*Note 1:* The methodology on how to determine the amount of data needed to meaningful calibrate a model is not addressed in this work, further reading on this subject can be found in [5].

### 2.2.1 k-fold Cross Validation

The $k$-fold cross-validation test divides the original sample into $k$ subsamples. The allocation of each element to each o subsamples is made randomly. Then, one of the k subsamples is selected to be the validation data set. The remaining $k-1$ subsamples are

used as calibration data set. The accuracy result obtained with this experiment is stored.

This experiment is repeated $k$ times, each time using a different subset as validation data set.

The $k$ accuracy results from each experiment are then averaged (or combined with other criteria) to produce a single accuracy estimation.

The $k$-fold cross-validation test uses all data elements for both calibration and validation. Each data element is used for validation once. In the general case, the number of subsamples $k$ to adopt is an undefined parameter.

As outlined by [13], re-sampling strategies have been commonly misused, often resulting in highly biased estimates of prediction.

### 2.2.2 Split sample test

The available data from reality should be split into two data sets to use in calibration and in validation. Each data set should be used in turn for calibration and for validation. Accordingly, one experiment would be conducted using the first part of the data for calibration and the second part of the data for validation, using a specified performance criterion. Another experiment would be conducted using the second part of the data for calibration and the first part of the data for validation, using the same performance criterion.

The results obtained from each experiment should be compared. The confirmation of the model would then be assessed by: (i) the similarity of the results from the two experiments and (ii) the adequacy of the validation outcomes with the corresponding ranges of accuracy. This means that the conceptual model may be validated if the validation outputs obtained with both experiments comply with the specified ranges of accuracy for the performance criteria used, and if these two validation outputs are similar. Otherwise, the conceptual model should be rejected.

For the author knowledge, there is no consensual definition on the ideal ratio used to split calibration and validation data sets. For instance in the study [21], it is suggested that when the available data is sufficiently long, two equal parts should be considered. In the study [13] several tests are conducted to address this same question, and the conclusions indicate ranges of ratios depending on the specific conditions of the model and of sample population.

### 2.2.3 Proxy-system test

The Proxy-system test is applicable to models designed to be transferable over systems. A transferable model is useful when there is low or any data available concerning the

system to be modeled.

#### Modelling a system with no data available

Let us consider a system *C* with any data available, and that we want to construct and validate a simulation model to simulate system *C*.

In this situation it is possible to develop a model with base on two other systems *A* and *B*, both with data available and with characteristics similar to system *C*, allowing the conjecture that these two systems are both representative of system *C*.

The model should be calibrated on system *A* and validated on system *B* (first experiment) and vice versa (second experiment). The model is *confirmed* (i.e. it is concluded that the model encompasses a basic level of credibility with regard to its ability to simulate system *C* adequately) only when the validation results from both experiments are similar and comply with the accuracy ranges specified.

#### Modelling a system with low data available

Let us consider a system *D*, with scarce data available (i.e. there are not sufficient data to perform a split as the one suggested in the split sample test). On one hand, the construction of a simulation model to simulate system *D* cannot be accomplished with the scarce data available, as it is not sufficient to perform calibration and validation. On the other hand, the scarce data available are the best knowledge about system *D*, and should be included within the simulation process.

In this case, the procedure to adopt under the calibration and validation processes is exactly the same described in the former section (Proxy-system test to model a system with any data available) with the attachment of a third experiment. The third experiment consists on using the scarce data set from system *D* for validation.

The model is accepted when the validation results from the three experiments comply with the accuracy ranges specified and are similar.

### 2.2.4 Differential split-sample test

The differential split-sample test is applicable to validate the sensitivity of the model to respond accurately on changes of a specific feature that integrates the model. This test may have several variants depending on the specific nature of the change to be simulated.

**Modeling a system sensitive to changes on a feature, with available data on the adjustable feature**

Let's consider a system $E$ to be simulated, for which there are sufficient data for calibration and validation processes. The simulation model of system $E$ is intended to be sensitive to changes on a specified feature $Y$ (the system $E$ includes a set of distinct features, one of which is feature $Y$).

In this case, the data available should be split into two parts according to the values observed in feature $Y$. Accordingly, the first data set would be composed with observations where feature $Y$ returned high values (e.g. considering the adjustable feature *birth rate of firms*, the first data set would include observations where high birth rates were observed). The second data set would include observations where feature $Y$ returned low values. This implies the exclusion of those observations where feature $Y$ returned moderate values.

- To test the model ability to reproduce the reality for high values of feature $Y$, the data set with low values of $Y$ should be applied for calibration, and the data set with high values should be used for validation. The model is confirmed if the validation results comply with the accuracy ranges specified. Otherwise, the model is rejected.
- To test the model adequacy to behave with low values of feature $Y$, the same reasoning is applied following a reverse order (the data set with high values is used for calibration, and the data set with low values of $Y$ is applied for validation).

**Modeling a system sensitive to changes on a feature, without significantly different data on the adjustable feature**

Let us consider a system $H$ to be simulated. The simulation model of $H$ is intended to be sensitive for changes on a feature $Y$ (system $H$ incorporates several features, being $Y$ one of them). Let us assume that, in the given record, segments with significantly different values of feature $Y$ cannot be identified (this situation may occur due to scarcity of data available or because the majority of observations relate to moderate values of $Y$).

In this case the model should be calibrated and validated on two substitute systems $F$ and $G$. The two substitute systems should (i) encompass sufficient and significantly different data on the adjustable feature (allowing the accomplishment of the differential split-sample test detailed in the former section), and (ii) be composed of similar characteristics to system $H$ (allowing the conjecture that systems $F$ and $G$ are representative of system $H$).

- To test the model adequacy to behave with high values of feature *Y*, two experiments should be conducted. The first experiment uses data from system F: the model is calibrated with data relative to low value of feature *Y*, and is further validated with data where high values of *Y* where observed. The second experiment uses data from system *G*, with similar reasoning as experiment 1. The model is accepted if the validation results comply with the specified accuracy ranges and if they are similar over the two experiments. The model is rejected otherwise.

- To test the model adequacy to reproduce low values of *Y*, the following two experiments should be conducted. The first experiment applies the data with high values of feature *Y* from system *F* for calibration, and the data with low values of *Y* from system *F* for validation. The second experiment applies data from system *G* with similar reasoning as the first experiment. The model acceptance or rejection is decided with base on the accuracy and similarity of the validation results obtained in both experiments.

*Note 2:* when using two substitute systems on differential split-sample test, the calibration and validation is done on each substitute system independently, which is different from the proxy-system test where a model is calibrated on one system and validated on the other.

*Note 3:* As reported in [21]: "A differential split-sample test can arise by default from a simple split-sample test if the only meaningful way of splitting an available record is such that the two segments exhibit markedly different conditions".

### 2.2.5 Proxy-system differential split-sample test

The Proxy-system differential split-sample test is applicable to models designed to be transferable both between systems and between features. The test may have different forms depending on the specific modeling task pursued.

Let us consider the system K with no available data records. System K includes several features, being one of them the feature Y. It is intended to test the model adequacy to replicate changes on feature Y within system K.

In this situation two other systems I and J, both representative of system K and with available and significantly different data on feature Y, should be chosen.

The data records of systems I and J are then split considering the values of feature Y. Accordingly, the data from system I is split into two parts, one referring to high values of Y, and the other concerning the low values of Y (observations referring to moderate

values of Y are not included). The data from system J is split following similar reasoning. At this point, four data sets are defined: high values of Y from system I; low values of Y from system I; high values of Y from system J; low values of Y from system J.

- To test the ability of the model to behave under high values of feature Y, two experiments are conducted. The first experiment uses the data from system I relative to low values of feature Y for calibration, and the data from system J relative to high values of feature Y for validation. The second experiment applies data from system J concerning low values of Y for calibration, and data from system I relative to high values of Y for validation. The model is accepted with base on the accuracy and similarity of the validation results obtained with the two experiments.

- The ability of the model to behave under low values of Y is tested with the following two experiments. The first experiment applies data from system I referring to high values of Y for calibration, and data from system J concerning low values of Y for validation. The second experiment applies the data set with high values of Y from system J for calibration, and the data with low values of Y from system I for validation. The decision on the model acceptance or rejection is made with base on the accuracy and similarity of the results obtained from the validation of the two experiments.

# Chapter 3

# Performance Criteria Overview

This chapter provides a bibliographic survey on performance criteria suitable to validate simulation models, using paired data samples. A performance criteria may be a measure, a metric, a statistical test or a procedure, whose application allows the assessment on the results quality obtained with a simulation model. For each approach revised a summary description is provided. Whenever possible, applicability examples and relevant critics in literature are referred.

The chapter is organized in seven sections. The first section encompasses error based methods to assess the divergence between two data samples. The second section includes measures of information theory, which allow the assessment of the amount of information contained in data sets. The third section includes two information criteria, Akaike Information Criteria and Bayesian Information Criteria (these measures are only applicable for models statistically defined, such as AR or MA). The fourth section includes parametric tests and measures. Section five provides a list of nonparametric tests that may be applicable to paired samples. Section six is focused on distance measures between two vectors, including a summary description of the Minkowski metric, its variants, short time series distance and the Dynamic Time Warping algorithm. The seventh section includes combined measures resulting from the assessment of the characteristics magnitude, phase and shape.

The notation used to describe each approach considers two data sets, both containing $N$ elements, being $i$ the index of an element: (i) $X = (x_1, x_2, ..., x_N)$ refers to the observations obtained from *reality*, (ii) and $\widehat{X} = \widehat{x}_1, \widehat{x}_2, ..., \widehat{x}_N$ encompasses the resulting estimated values from a simulation model.

## 3.1 Error-based measures

This section summarizes measures that can be used to validate simulation models based on the error obtained considering the simulated outputs and the respective observed data. The error between two homologous elements from each data set $e_i$ is defined in equation (3.1) as the difference between the $i^{th}$ element from observed data set $x_i$ and the homologous $i^{th}$ element from simulated data set $\widehat{x}_i$. Note that $e_i$ is on the same scale as the data sets.

$$e_i = x_i - \widehat{x}_i \tag{3.1}$$

The error can be perceived as the distance between two ordinates (one simulated and one observed) relating to a specific abscissa (e.g. time line), i.e. the divergence observed between two data sets in a specific moment.

According to [18, 19], there are four types of error based measures: (i) *scale-dependent* measures, (ii) *percentage-error* measures, (iii) *relative-error* measures and (iv) *scale-free error* measures. The next sections include the most widely used error-based metrics organized within these four main groups. An additional group is considered which includes the Theil's measures on accuracy and on quality of the forecasts.

### 3.1.1 Scale-dependent measures

As explained in [18], *scale-dependent* measures cannot be used to compare accuracy results across case studies with different data units, as the values obtained are scale dependent. This is the main drawback of these measures. Nevertheless, *scale-dependent* measures are powerful performance criteria to assess the similarity of data sets provided they are within the same unitary system. Moreover, the measures described in this section based many other measures.

#### Mean Error - $ME$

The Mean Error $ME$, defined in equation (3.2), provides the average error obtained, considering the exact values of $e_i$.

$$ME = \frac{1}{N} \cdot \sum_{i=1}^{N} e_i \tag{3.2}$$

$ME$ can return both positive and negative values. A better fit of the estimated data return values of $ME$ close to zero. The $ME$ measure does not prevent negative and positive errors from offsetting each other, which is a drawback of this measure.

**Mean Absolute Error -** $MAE$

The Mean Absolute Error $MAE$ is defined as the mean of the absolute values of $e_i$, as presented in equation (3.3).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|e_i| \tag{3.3}$$

This measure is similar to $ME$, but it considers the absolute value of the error instead. The use of absolute values prevents negative and positive errors from offsetting each other, from where the values obtained with $MAE$ are always positive. A better fit of the estimated data set is linked with lower values of $MAE$. A comprehensive study comparing $MAE$ and $RMSE$ (see section 3.1.1) can be found in [56], where $MAE$ is found to be more natural and unambiguous than $RMSE$.

**Mean Square Error -** $MSE$

The Mean Square Error $MSE$ is the average of the squared errors, as defined in equation (3.4).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}e_i^2 \tag{3.4}$$

The inclusion of the square error in the calculus of $MSE$ induces three main aspects in the values obtained: (i) it returns always positive values, and a better fit of the estimated set is associated with values of $MSE$ close to zero, (ii) this measure may not be interpreted in the same units as the data sets under analysis (for errors lower than $1$ the corresponding $MSE$ is lower than the observed errors, for errors higher than $1$ the $MSE$ will return higher values), and (iii) this measure prevents the offsetting of negative and positive errors.

**Root Mean Square Error -** $RMSE$

The Root Mean Square Error $RMSE$ is the square root of the $MSE$, as defined in equation (3.5).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}e_i^2} \tag{3.5}$$

$RMSE$ returns positive values and a better fit of the model is linked with values of $RMSE$ close to zero. This measure prevents offsetting of negative and positive error

values. The use of the square root to the sum of the square errors makes this measure tricky to interpret, in the sense that the wrong idea of "$RMSE$ returns values in the same units as the data sets" easily occurs. As cautioned in the study [56], $RMSE$ is a frequently misinterpreted measure of average error because it is the result of the interaction of 3 characteristics of a set of errors: (i) the variability within the distribution of error magnitudes, (ii) the square root of the number of errors ($N^{0.5}$) and (iii) the average-error magnitude. The study [56] also suggests that $MAE$ (see section 3.1.1) should be adopted instead of $RMSE$ as it is a more intuitive measure of average error assessment. The study [56] concludes that "it seems to us that there is no clear interpretation of RMSE or related measures".

Another study [55] explores the $RMSE$ alongside with the correlation coefficient (see section 3.4.1) and the coefficient of determination (see section 3.4.2). This study concludes that $RMSE$ is a superior criterion to compare simulated and observed data sets than correlation coefficient or coefficient of determination.

### 3.1.2 Percentage-error measures

All metrics constructed based on *percentage-error* have the advantage of being scale independent, as explained in [18]. Thus, these measures may be compared across different case studies. The percentage error between two homologous elements from each data set $PE_i$ is defined as the ratio between $e_i$ and the respective observed element $x_i$, as defined in equation (3.6).

$$PE_i = \frac{e_i}{x_i} \tag{3.6}$$

The main disadvantage of *percentage-error* measures, also referred in [18], is that they return undefined values for historic data elements with value of zero.

#### Mean Percentage Error - $MPE$

The Mean Percentage Error $MPE$ is calculated as the mean of the percentage errors $PE_i$, detailed in (3.7).

$$MPE = \frac{1}{N}\sum_{i=1}^{N} PE_i \tag{3.7}$$

$MPE$ may return both positive or negative values. A better fit of the estimated data set is associated with values of $MPE$ close to zero. This measure has the drawback of

offsetting positive and negative values of $PE$.

**Mean Absolute Percentage Error -** $MAPE$

The Mean Absolute Percentage Error $MAPE$ is the mean of the absolute values of $PE_i$, as detailed in (3.8).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}|PE_i| \tag{3.8}$$

The values obtained with $MAPE$ are always positive, and a better fit of the estimated data set is identified with lower values of $MAPE$. This measure has the advantage of not offsetting positive and negative values of $PE$. In [4] it is stated that a disadvantage of $MAPE$ is that it is relevant only for ratio-scaled data (i.e., data with a meaningful zero).

### 3.1.3 Relative-error measures

*Relative-error* measures are summarily described in this section as a matter of consistency with this bibliographic survey, but they are not adequate to compare estimated results with the respective historic data set. These methods are designed to compare the errors obtained with two different models: (i) a model that is new, and (ii) a second model that is accepted in literature as good, used as a benchmark model. In this work, the main focus is directed to performance criteria adequate to compare the fitness of a single model. For this reason, the *relative-error* measures are not tested in further sections.

*Relative error* measures represent an alternative to *percentage-error* measures as they return scale independent values. Lets consider the error obtained with the new model $e_i$ and the error obtained with the benchmark model $e_i^b$ for an $i^{th}$ homologous element. The ratio between $e_i$ and $e_i^b$ is then calculated according to equation (3.9).

$$r_i = e_i/e_i^b \tag{3.9}$$

A *relative-error* measure is then easily obtained using the ratio $r_i$ instead of the error $e_i$ for any *scale-dependent* measure. For example, the calculus of the Relative Mean Absolute Error $RMAE$ measure would be made as specified in equation (3.10).

$$RMAE = \frac{1}{N}\sum_{i=1}^{N}|r_i| \tag{3.10}$$

These methods are suggested in [4], but they are rarely used in practice. The con-

clusions taken from applying a *relative-error* measure or comparing the respective *scale-dependent* measure to both models is equivalent. Moreover, these methods are not applicable when the error obtained from the benchmarking model is zero, as in that case the value of $r_i$ would be undefined [18].

### 3.1.4 Scale-free error measures

According to Hyndman [18], the main advantage of *scale-free error* measures is that they provide more accurate results than the methods reviewed in the former three sections when the data sets to be assessed are non stationary, meaning that the data evolve according to a pattern such as trend or seasonality. This does not inculcate any inadequacy of the methods reviewed in the former three sections to the assessment of non-stationary data. Important to note that the *scale-free error* measures are recent, and there is still few comparative literature.

*Scale-free error* measures are based on *scale-dependent* measures, but they use a scaled error $q_i$ instead of error $e_i$. The scaled error is calculated as described in equation (3.11).

$$q_i = \frac{e_i}{\frac{1}{N-1}\sum_{j=2}^{N}|x_j - x_{j-1}|} \tag{3.11}$$

**Mean Absolute Scaled Error -** $MASE$

For the author knowledge, it was possible to find only one *scale-free error measure* application in literature, referring to Mean Absolute Scaled Error $MASE$, although it is suggested that the reasoning of replace a scaled error $q_i$ by the classic error $e_i$ may be extended to other measures. The $MASE$ is calculated as presented in equation (3.12), which is the same as using equation (3.3) replacing $e_i$ by $q_i$. This method is applied in [18].

$$MASE = \frac{1}{N}\sum_{i=1}^{N}|q_i| \tag{3.12}$$

Hyndman [18, 19] claims that $MASE$ is the only available accuracy measurement that can be used in all forecasting situations and for all types of series. Note that this claim is based on a comparison with error-based measures only, and is justified with the following advantages identified for this method: it is scale free and therefore the values

obtained can be compared across models with different unitary systems, when applied to non-stationary data the results obtained are more accurate than with other error-based measures, it does not incur in undefined elements, it prevents negative and positive errors from offsetting each other.

### 3.1.5 Theil's measures

**Theil's measure of forecast accuracy -** $U_1$

$U_1$ was proposed by Theil in 1966 [51] as a measure of forecast accuracy, as specified in (3.13).

$$U_1 = \frac{\left[\frac{1}{N}\sum_{i=1}^{N}(e_i)^2\right]^{0.5}}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i)^2\right]^{0.5} + \left[\frac{1}{N}\sum_{i=1}^{N}(\widehat{x}_i)^2\right]^{0.5}} \tag{3.13}$$

When $U_1 = 0$, it means that the estimation is completely coincident with the observations ($x_i = \widehat{x}_i, \forall i$), indicating a perfect forecast. The case $U_1 = 1$ indicates the maximum inequality (when there is negative proportionality over the two data sets or when one of the data sets is identically to zero) [8, 51].

Bliemel [8] analyzed both measures proposed by Theil, $U_1$ and $U_2$ (see next section), and concluded that $U_1$ is only informative to assess forecast accuracy when applied to the absolute values of the errors.

**Theil's measure of forecast quality -** $U_2$

The measure $U_2$ was proposed by Theil in 1965 [50] to assess the quality of forecasts. This measure is defined in (3.14).

$$U_2 = \frac{\left[\frac{1}{N}\sum_{i=1}^{N}(e_i)^2\right]^{0.5}}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i)^2\right]^{0.5}} \tag{3.14}$$

The result $U_2 = 0$ indicates a perfect forecast, meaning that both data sets are coincident ($x_i = \widehat{x}_i, \forall i$). The meaning of the case $U_2 = 1$ is clarified by [8], stating that this is observed "when the prediction method is naive no-change extrapolation or when it leads

to the same standard deviation of forecast error as that method".

The comparative study performed by Bliemel [8] over the two Theil's measures, $U_1$ and $U_2$, concludes that $U_2$ provides more meaningful information on the accuracy of the estimations under assessment, suggesting that $U_2$ should be preferably used than $U_1$.

As clarified in [33], the statistic $U_2$ is normally applied without the formal definition of hypothesis. Nevertheless, it can be applied as the test statistic of parametric test when both data sets come from bivariate normal populations. The distribution $F$ is applicable in that case. Further reading on this topic can be found in [33].

## 3.2 Information Theory $IT$ based measures

The Information Theory $IT$ field was firstly developed by Shannon [46], with the definition of the concepts of entropy and mutual information. The $IT$ concepts have been applied in several research fields since then, such as in biology [1], simulation of agent based models [52] and computational cybernetics [31].

The measures used in $IT$ are mathematical quantities that record the amount of information contained within a data set. In this section the measures entropy and Kullback-Leibler Divergence are reviewed as they were proposed by Akaike [2] to be applied to validate simulation models. More recently, normalized information theory measures have been suggested to assess the goodness of fit of simulation models, see [53]. These normalized measures are also reviewed.

The observed data set $X$ is assumed to have, at least, some randomness (otherwise it would be a deterministic phenomena, and the respective simulation model would return perfect estimations). Therefore lets consider the random variable $X$, with $P = (p_1, p_2, ..., p_N)$ being the probabilities corresponding to each discrete observation $X = (x_1, x_2, ..., x_N)$, and $\sum_{i=1}^{N} p_i = 1$.

### 3.2.1 Entropy

Entropy was firstly defined by Shannon in 1948 [46]. According to Shannon [46], entropy is the quantity of information linked to the probability of occurrence of a certain event: (i) entropy is null for events whose output is completely known at start and (ii) entropy is higher the less predictable an event is. Entropy was proposed as a performance criterion by Akaike in 1974 [2] under the context of time series estimations, and was studied in [36] under the context of Markov models.

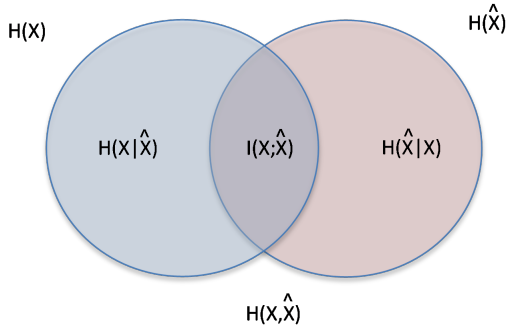


Figure 3.1: Schematic representation of Information Theory quantities.

The general mathematical definition of entropy was proposed by Renyi, the Renyi's Entropy - $H_{r\alpha}$, which is presented in (3.15), being $\alpha$ a real parameter subject to the restrictions specified in (3.15), and therefore the entropy firstly described by Shannon $H_s$ is a particular case of the general entropy described by Renyi $H_{r\alpha}$.

$$\begin{aligned} H_{r\alpha}(X) &= \tfrac{1}{1-\alpha} log\left(\sum_{i=1}^{N} p_i^{\alpha}\right) \\ \text{Subject to: } \alpha &> 0 \\ \alpha &\neq 1 \end{aligned} \tag{3.15}$$

Figure 3.2.1 provides a schematic visualization on information theory quantities of entropy, joint entropy and mutual information.

In Figure 3.2.1 the entropy of the observed data set $H(X)$ is represented as the blue circle, the entropy of the estimated data set $H(\widehat{X})$ is represented as the red circle. The joint entropy over the two data sets $H(X, \widehat{X})$ is represented as the union $H(X) \cup H(\widehat{X})$. The mutual information (see section 3.2.2) is represented as the intersection $H(X) \cap H(\widehat{X})$.

### Quadratic entropy of Renyi

In the case of $\alpha = 2$ the expression (3.15) leads to the quadratic entropy, due to the quadratic form of the probability, as defined in (3.16).

$$H_{r2}(X) = -log\left(\sum_{i=1}^{N} p_i^2\right) \tag{3.16}$$

The quadratic entropy of Renyi has the advantage of being easier to calculate through the application of Gaussian convolution [7].

**Entropy described by Shannon -** $H_s$

The entropy described by Shannon $H_s$ is the weighted sum of logarithms by probabilities. It represents the average amount of information included within a single observation of $X$. Accordingly, the Shannon's entropy $H_s(X)$ is defined in (3.17).

$$\begin{aligned} H_s(X) &= \sum_{i=1}^{N} p_i \cdot log\left(\frac{1}{p_i}\right) \\ \text{Subject to: } \sum_{i=1}^{N} p_i &= 1 \\ p_i &\geq 0 \\ 0 \cdot log_2(0) &= 0 \end{aligned} \tag{3.17}$$

The use of entropy as a validation criterion is however tricky. The direct comparison between the entropy of the historic data set with the entropy of the simulated data set do not allow to conclude on the adequacy of the results obtained. As it can be perceived in image 3.2.1, it is possible to have two independent data sets returning similar values for entropy $H(X)$ and $H(\widehat{X})$, as in that case the mutual information would be zero.

The use of entropy as validation criteria, in the light of this work, should be done considering the two data sets simultaneously, using the joint entropy over the two data sets $H(X, \widehat{X})$. A perfect fit would be obtained for a value of joint entropy equal to the value of the observed entropy ($H(X, \widehat{X}) = H(X)$). The join entropy is calculated as described in (3.15), using the attached observations of $X$ and $\widehat{X}$ instead of using the observations $X$ separately.

### 3.2.2 Kullback-Leibler Divergence - $D_{KL}$

The Kullback-Leibler Divergence $D_{KL}$ is a mathematic quantity which measures dissimilarity between two different probability density functions $p(x)$ and $q(\widehat{x})$ ([39] pp. 16). This measure is defined as follows:

$$D_{KL}(p\|q) = \sum_{x} p(x) log \frac{p(x)}{q(\widehat{x})} \tag{3.18}$$

The divergence $D_{KL}$ may be interpreted as the distance between the two probability density functions, however it does not obey the distance mathematical postulates ([39] pp. 16). The divergence $D_{KL}$ is often used to validate simulation models, as it is suggested

in [29]. An example is provided by [20], with the validation of the results obtained with a simulated annealing model.

**Mutual Information -** $I$

Mutual information $I$ is a special case of the divergence $D_{KL}$: when the two probability density functions under assessment are (i) the joint probability density function of $p(x, \widehat{x})$ and (ii) the product of the marginals of $p(x)q(\widehat{x})$ ([39] pp. 18). Mutual information $I$ is defined as follows.

$$I(X, \widehat{X}) = D_{KL}(p(X, \widehat{X}) \| p(X)q(\widehat{X})) \tag{3.19}$$

In the scope of this thesis, the mutual information measure may be useful when applied to the two data sets under assessment. This measure would allow the measurement of the divergence between the distributions. The measure $I(X, \widehat{X})$ returns values in the range $[0, min(H(X), H(\widehat{X}))]$. A perfect fit of the estimated data set would return the maximum value of $I(X, \widehat{X})$, as in that situation $I(X, Y) = H(X) = H(\widehat{X})$. A bad estimated data set would return $I(X, \widehat{X})$ values close to zero.

The study [28] compares mutual information and correlation measures, clarifying that although both measure dependence over data sets, mutual information measures a general dependence, while the correlation function measures a linear dependence. Thus, mutual information is considered a better quantity than the correlation function to measure the dependence over two data sets.

### 3.2.3 Normalized Information Theory measures

The normalized $IT$ measures were constructed with base on the $IT$ measures defined in the previous two sections. These measures are applied in [53]. A summary of normalized $IT$ measures is provided in (3.20) concerning Normalized Mutual Information $NMI_{criteria}$, and in (3.21) concerning Normalized Information Distance Measures $Id_{criteria}$.

$$
\begin{aligned}
NMI_{joint} &= \frac{I(X,\widehat{X})}{H(X,\widehat{X})} \\
NMI_{max} &= \frac{I(X,\widehat{X})}{max(H(X),H(\widehat{X}))} \\
NMI_{min} &= \frac{I(X,\widehat{X})}{min(H(X),H(\widehat{X}))} \\
NMI_{sum} &= \frac{2 \cdot I(X,\widehat{X})}{H(X)+H(\widehat{X})} \\
NMI_{sqrt} &= \frac{I(X,\widehat{X})}{\sqrt{(H(X) \cdot H(\widehat{X}))}}
\end{aligned}
\tag{3.20}
$$

All Normalized Mutual Information detailed in 3.20 return values within the range $[0, 1]$. Values close to zero indicate a bad estimation, whilst values close to one indicate a good estimation.

$$
\begin{aligned}
Id_{joint} &= 1 - \frac{I(X,\widehat{X})}{H(X,\widehat{X})} \\
Id_{max} &= 1 - \frac{I(X,\widehat{X})}{max(H(X),H(\widehat{X}))} \\
Id_{min} &= 1 - \frac{I(X,\widehat{X})}{min(H(X),H(\widehat{X}))} \\
Id_{sum} &= 1 - \frac{2 \cdot I(X,\widehat{X})}{H(X)+H(\widehat{X})} \\
Id_{sqrt} &= 1 - \frac{I(X,\widehat{X})}{\sqrt{(H(X) \cdot H(\widehat{X}))}}
\end{aligned}
\tag{3.21}
$$

The Normalized Information Distances Measures return values within the range $[0, 1]$, with zero indicating a good estimation and one indicating a bad estimation.

## 3.3 Validation using Information Criteria - IC

The validation of simulation models may be performed using Information Criteria, when the model to be validated is statistically defined, such as a linear model or and auto-regressive moving average $ARMA$ model. When the model to be validated is not definable statistically, meaning that the validation is to be performed using only the data sets (that is the case of System Dynamics models), information criteria are not applicable.

In [23], several information criteria are presented, including: (i) Akaike Information Criterion $AIC$, (ii) Bayesian Information Criterion $BIC$ (also known as Schwarz Information Criterion $SIC$), Deviance Information Criterion $DIC$, (iv) Extended Information Criterion $EIC$, (v) Focused Information Criterion $FIC$, (vi) Generalized Information Criterion $GIC$, (vii) Network Information Criterion $NIC$, and (viii) Takeuchi's Information Criterion $TIC$. According to [23], the majority of the criteria included in this list are modifications or generalizations of $AIC$ and $BIC$, which are the two more commonly used.

### 3.3.1 Akaike Information Criterion - $AIC$

A widely used model-selection criterion is Akaike's Information Criterion (AIC) (see [3, 30, 23]). AIC is an asymptotically unbiased estimator of the expected Kullback-Leibler divergence $D_{KL}$ (see equation 3.18) between two probability density functions [38, 23, 2], normally applied to assess the quality of the results obtained with different models when historic data is available. Thus, the historic data is used as the reference, and the results obtained with AIC allow to measure the divergence between the (i) probability density functions obtained with the simulated results and (ii) the probability density function of the historic data. Therefore, the estimated model will be closer to reality, the lesser the distance between the two probability density functions considered. Thus, smaller values of $AIC$ indicate a better fit of the model.

The $AIC$ for a given model is a function of its maximized value of the likelihood function for the estimated model ($L$) and the number of estimable parameters ($K$), as defined in equation (3.22).

$$AIC = 2K - 2\ln(L) \tag{3.22}$$

Another similar criteria, the Minimum Information Theoretic Criterion Estimate - $MAICE$, was also proposed by Akaike in [2], as an improvement of the $AIC$. Accordingly, the $MAICE$ eliminates the "need of the subjective judgment required in the

hypothesis testing procedure for the decision on the levels of significance". In this work, the $AIC$ is adopted as it is much widely used.

### 3.3.2 Bayesian information criterion - $BIC$

The Bayesian information criterion $BIC$, or Schwarz criterion (also $SBC$, $SIC$ and $SBIC$) is normally applied to choose the best simulation model from a set of models. It assesses the quality of adjustment of a simulation model. $BIC$ is a widely applied measure, normally alongside with $AIC$, see ([30] pp. 173, [23]). A detailed description and critic on $BIC$ can be found in [54] and in [23]. $BIC$ is defined in equation (3.23), where $r$ is the number of degrees of freedom remaining after fitting the model. Smaller values of $BIC$ indicate a better fit of the model.

$$BIC = L^2 - r\ln(N) \tag{3.23}$$

In the study [23], the criteria $AIC$ and $BIC$ are examined alongside, and the use of these two criteria simultaneously to base the assessment of the quality of models results is advised, as these criteria approximate two different target quantities.

## 3.4 Parametric tests

It is frequent to apply a *Parametric* test when there is a prior assumption that a sample came from a distribution of a particular family (examples of distribution families are Gaussian, Binomial, Exponential, Gamma, Beta or Weibull). The parametric tests main aim is to find a statistically significant estimation of a specific parameter, using an hypothesis approach. In this section the Pearson's coefficient of correlation and the coefficient of determination are reviewed, as these two parametric tests are widely used to assess simulation models. The use of correlation-based measures is criticized in [25, 55]. These studies highlight that correlation-based measures are oversensitive to extreme values (outliers) and insensitive to additive and proportional differences between model predictions and observations. These limitations can induce in wrong acceptance of simulation models, as documented in [55].

Lets consider the mean of the historic data set and the mean of the estimated data set as defined in (3.24).

$$\begin{aligned} \overline{X} &= \tfrac{1}{N}\sum_{i=1}^{N} x_i \\ \overline{\widehat{X}} &= \tfrac{1}{N}\sum_{i=1}^{N} \widehat{x}_i \end{aligned} \tag{3.24}$$

### 3.4.1 Coefficient of correlation - $r$

The coefficient of correlation, or Pearson's Product-Moment Correlation Coefficient $r$, represents the degree of linear association between two variables and is mathematically defined in (3.25).

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(\widehat{x}_i - \overline{\widehat{X}})}{\left[\sum_{i=1}^{N}(x_i - \overline{X})^2\right]^{0.5}\left[\sum_{i=1}^{N}(\widehat{x}_i - \overline{\widehat{X}})^2\right]^{0.5}} \tag{3.25}$$

The coefficient $r$ may assume values between the range $[-1, 1]$, and this value is absolute and non-dimensional. The interpretation of this coefficient is detailed in [49]. Accordingly, (i) a correlation coefficient of zero indicates that no association exists between the measured variables, (ii) a positive correlation coefficient indicates that an increase in the first variable would correspond to an increase in the second variable, and (iii) a negative correlation indicates that whereas one variable increases, the second variable decreases.

As Taylor [49] explains, a statistically significant $r$ coefficient can only indicate that the observed sample data provides evidence to reject the null hypothesis that the population correlation coefficient parameter $r$ is zero. The rejection of the null hypothesis allow to conclude that the correlation coefficient of the population is not equal to zero. Under the context of validation of simulation models, it is desirable a value of $r$ statically significant and positive, therefore the hypothesis test should be formulated as specified in (3.26). The application of this test can only be made under the following assumptions considering both samples: the samples are random, quantitative, normally distributed, linearly related and have the same variance (homoscedasticity).

$$\begin{aligned} H_0 &: r = 0 \\ H_1 &: r > 0 \end{aligned} \tag{3.26}$$

The test proceeds with the calculation of the test statistic value as defined in equation (3.27).

$$TS = r\sqrt{\frac{N-2}{1-r^2}} \mapsto T_{N-2} \tag{3.27}$$

The $TS$ value obtained is compared with the critical region $CR = ]C, +\infty[$, where $C$ is to be chosen using Student's t-distribution tables, according to the desired level of significance. A comprehensive description of this test is provided by Sprent ([47], pp. 163), and the Student's t-distribution tables can be found in [11]. The proposal of this approach to validate simulation models can be found in several studies, such as [44, 43].

### 3.4.2 Coefficient of determination - $R^2$

The coefficient of determination $R^2$ is a widely used indicator to assess the quality of simulation models. $R^2$ is the square of the Pearson s Product-Moment Correlation Coefficient [25, 55], and it provides a measure on the global adequacy of the model, by reporting the proportion of variability in a historic data set that is reproduced in the simulated data set (note that proportion is not a percentage). The same assumptions made in section 3.4.1 are adopted here. $R^2$ is mathematically defined in equation (3.28).

$$R^2 = r^2 = \left\{ \frac{\sum_{i=1}^{N}(x_i - \overline{X})(\widehat{x}_i - \overline{\widehat{X}})}{\left[\sum_{i=1}^{N}(x_i - \overline{X})^2\right]^{0.5}\left[\sum_{i=1}^{N}(\widehat{x}_i - \overline{\widehat{X}})^2\right]^{0.5}} \right\}^2 \tag{3.28}$$

The values obtained with $R^2$ are always positive and included within the range $[0, 1]$. The closest $R^2$ is to the value 1, the best the model reproduces reality.

According to [25], $R^2$ is limited in that it standardizes for differences between the observed and predicted means and variances since it only evaluates linear relationships between the variables. In [25] it is also mentioned that $R^2$ is insensitive to additive and proportional differences between simulated and observed homologous elements, and that large values of $R^2$ can be obtained even when the magnitude between the observed-

simulated values is considerably different.

When assessing statistic models such as the linear regression, the $R^2$ value obtained may be used as the test statistic for a hypothesis test, using the F-distribution. This parametric test can only be made for statistic models because the information on the number of degrees of freedom is needed [17]. Therefore, for a general model with two data sets to be compared, this hypothesis test is not applicable and the goodness of fit is evaluated considering ony the value returned by $R^2$.

### 3.4.3 Cross correlation - $\rho(n)$

The cross correlation $\rho(n)$, also known as sliding dot product, describes the normalized cross covariance function between two data sets. Cross correlation may be used to find the phase lag between the two data sets, being the phase lag to be found $n = 0, 1, \cdots, N-1$. The same assumptions made in sections 3.4.1 and 3.4.2 are adopted here. The formulation of $\rho(n)$ is presented in (3.29) accordingly to the studies [44, 43].

$$\rho(n) = \frac{(N-n)\sum_{i=1}^{N-n} x_i \widehat{x}_{i+n} - \sum_{i=1}^{N-n} x_i \sum_{i=1}^{N-n} \widehat{x}_{i+n}}{\left[(N-n)\sum_{i=1}^{N-n} x_i^2 - \left(\sum_{i=1}^{N-n} x_i\right)^2\right]^{0.5} \left[(N-n)\sum_{i=1}^{N-n} \widehat{x}_{i+n}^2 - \left(\sum_{i=1}^{N-n} \widehat{x}_{i+n}\right)^2\right]^{0.5}} \tag{3.29}$$

The $\rho(n)$ is calculated for all possible values of $n = 0, 1, \cdots, N-1$. The maximum $\rho(n)$ found is chosen and the correspondent $n$ is the estimated lag between the two data sets. In order to clarify the notation, the estimated lag is here denoted by $n_{lag}$. This procedure is normally applied in signal processing, pattern recognition and cryptanalysis.

## 3.5 Nonparametric tests

The *Nonparametric* tests, also referred as *distribution free methods*, are adequate when there is no evidence that the samples arrive from a specific family of distributions.

In the light of the present work, it is interesting to revise non-parametric tests which

may allow to conclude whether an estimated data set is a good fit of the observed data set. The non-parametric tests adequate to this purpose are those who consider paired observations, also known as two matched samples, allowing for the method to assess the similarity of two samples according to the ordered values of homologous elements. According to the work developed by Conover [11], the appropriate tests to consider are presented in Table (3.1).

| **Hypothesis Test** | **Nominal data** | **Ordinal data** | **Interval data** |
|---|---|---|---|
| **Means and medians** | - McNemar Test | - Sign Test for means | - Wilcoxon Test;<br>- van der Waerden Test;<br>- Randomization Test |
| **Confidence intervals** for differences between means | - Confidence interval for $p$ | - Confidence interval for $x_i$ | - Confidence interval for differences |
| **Regression slope** | (none) | (none) | - Testing the slope |
| **Independence** | - Chi-square test;<br>- Fisher's exact test | - Sign test for independence;<br>- Spearman's rho;<br>- Kendall's tau | (none) |

Table 3.1: Nonparametric tests for paired observations (adapted from [11]).

The table 3.1 shows different tests to apply depending on the kind of data that is treated. In this work the test referring to ordinal data are going to revised. All the methods are clearly detailed in [11].

### 3.5.1 Sign test

The data used to perform the sign test is organized in paired samples, and it can be perceived as a bivariate random sample $(X, \widehat{X})$ with $N$ pairs of observations. Each pair $(x_i, \widehat{x}_i)$ is compared: the pair is classified as "$+$" if $x_i < \widehat{x}_i$, as "$-$" if $x_i > \widehat{x}_i$ or as "0" if $x_i = \widehat{x}_i$. The probability of a pair to be classified with "$+$" is $P(+)$, and the probability of a pair being classified with "$-$" is $P(-)$.

The test is constructed under the following assumptions: (i) the bivariate random observations $(x_1, \widehat{x}_1)$, $(x_2, \widehat{x}_2)$, ..., $(x_N, \widehat{x}_N)$ are independent; (ii) the pairs are internally consistent, meaning that the probability of a pair to be classified with "$+$" or "$-$" is the same for all pairs.

This test may be applied following two tails or one tail approach. Here, the two tails approach is adopted as the test is made to assess if "$X$ and $\widehat{X}$ have the same location parameter median". Accordingly, the hypothesis to formulate this test is defined in equation (3.30).

$$\begin{aligned} H_0 &: P(+) = P(-) \\ H_1 &: P(+) \neq P(-) \end{aligned} \tag{3.30}$$

The test statistic $TS$ used in the sign test is the number of pairs classified with "+" (the number of pairs with $x_i < \widehat{x}_i$). The number of pairs classified with a tie ("0") is removed from the original bivariate sample, so that the sample size used onwards is defined as $N^{'} = N -$ number of pairs with tie. The critical region is defined as presented in (3.31).

$$CR = [0, c_1[\cup]c_2, N^{'}] \tag{3.31}$$

The values $c_1$ and $c_2$ are found in the tables of the binomial distribution (see [11], pp. 433) to a specified significance level $\alpha$ for values of $N^{'}$ lower than 20.

For $N^{'}$ values higher than 20, $c_1$ and $c_2$ are calculated as defined in (3.32), being $\omega_{\alpha/2}$ obtained from Normal Distribution tables (see [11], pp. 124). For the common value $\alpha = 0.05$, $\omega_{\alpha/2} = -1.96$.

$$\begin{aligned} c_1 &= 0.5(N^{'} + \omega_{\alpha/2}\sqrt{N^{'}}) \\ c_2 &= N^{'} - c_1 \end{aligned} \tag{3.32}$$

If $TS \in CR$, the $H_0$ is rejected. If $TS \notin CR$, then the $H_0$ is accepted and there is statistical evidence that both data sets, the estimated and the observed, have identical medians with a confidence coefficient of $1 - \alpha$.

### 3.5.2 Spearman's rank correlation coefficient - $r_S$

Spearman's rank correlation coefficient $r_S$ is a non-parametric measure of statistical linear dependence between two variables. The Spearman correlation is normally used when the assumptions of the Pearson correlation are violated, and it is applicable to ordinal and quantitative variables. The first step is to assign a rank order to each element of each data set, using the average rank for draws. Therefore, two distinct rank sets are established: $R_1^X, R_2^X, \cdots, R_N^X$ is the rank set related to the historic data set, and $R_1^{\widehat{X}}, R_2^{\widehat{X}}, ..., R_N^{\widehat{X}}$ is the rank set related to the estimated data set. The difference between all homologous elements in the two rank sets is performed, originating the differences set and being the $i^{th}$ element defined as $D_i = R_i^X - R_i^{\widehat{X}}$. The Spearman rank $r_S$ can be calculated as

detailed in equation (3.33).

$$r_S = 1 - 6 \sum_{i=1}^{N} \frac{D_i^2}{N(N^2-1)} \tag{3.33}$$

Similarly to $r$, $r_S$ also vary between the range $[-1, 1]$, and is an absolute and non-dimensional value. A $r_S$ value of zero indicates no correlation between the variables, a positive value of $r_S$ indicates that an increase in $X$ corresponds to an increase in $\widehat{X}$, a negative value of $r_S$ indicates that an increase in $X$ is associated with a decrease in $\widehat{X}$. Once pursuing the objective of validate the results obtained with a simulation model, it is desirable to obtain statistically significance for positive values of $r_S$. This is made with the specification of an hypothesis test, formulated as presented in equation (3.34).

$$\begin{aligned} H_0 &: r_S = 0 \\ H_1 &: r_S > 0 \end{aligned} \tag{3.34}$$

The test statistic for this test is the value $r_S$. The critical region is defined as $CR = ]C, +\infty[$, where $C$ is tabulated according to the size of the samples $N$, and according to the level of significance to be tested. These tables may be found in [57]. If $r_S \in CR$, the $H_0$ is rejected and there is statistical evidence to assume that there is a positive correlation between the two data sets. If $r_S \notin CR$, then the $H_0$ is not rejected, and no conclusion may be formulated.

### 3.5.3 Kendall's tau - $\tau$

The Kendall's tau test is applicable to a bivariate random variable $(X, \widehat{X})$, composed of $N$ pairs $(x_i, \widehat{x_i})$.

The test starts with the comparison of each pair $(x_i, \widehat{x_i})$ with all the remainder pairs. Each comparison between two pairs returns a classification of *concordant*, *discordant* or *tie*, as specified next. The total number of comparisons to be made is $N(N-1)/2$.

- A comparison between two pairs $(x_i, \widehat{x_i})$ and $(x_j, \widehat{x_j})$ is concordant when $x_i > x_j \wedge \widehat{x_i} > \widehat{x_j}$ or $x_i < x_j \wedge \widehat{x_i} < \widehat{x_j}$.
- A comparison between two pairs $(x_i, \widehat{x_i})$ and $(x_j, \widehat{x_j})$ is discordant when $x_i > x_j \wedge \widehat{x_i} < \widehat{x_j}$ or $x_i < x_j \wedge \widehat{x_i} > \widehat{x_j}$.

- A comparison between two pairs $(x_i, \widehat{x_i})$ and $(x_j, \widehat{x_j})$ is a tie when $x_i = x_j$ or $\widehat{x_i} = \widehat{x_j}$.

Lets consider the total number of concordant comparisons as $N_c$, the total number of discordant comparisons as $N_d$. The test statistic is then formulated, here denoted as Kendall's $\tau$, defined in (3.35). When $N_c = N$, $\tau = 1$, and when $N_d = N$, $\tau = -1$.

$$\tau = \frac{N_c - N_d}{N(N-1)/2} \tag{3.35}$$

Kendall's $\tau$ may be used as a test statistic, with the hypothesis specified in (3.36). In this case, the test statistic is $TS = N_c - N_d$.

$$\begin{aligned} H_0 &: \tau = 0 \\ H_1 &: \tau > 0 \end{aligned} \tag{3.36}$$

The critical region for $TS$ is defined as $]C, +\infty[$. The value $C$ is tabulated according with the size of the samples $N$, and with the level of significance to be tested (see [11] pp. 458).

$H_0$ is rejected when $TS \in CR$, in this case there is statistical evidence to conclude for a positive correlation between $X$ and $\widehat{X}$. When $TS \notin CR$, the $H_0$ is not rejected, and any conclusion is made.

### 3.5.4 Wilcoxon Signed Rank Test

This test is performed considering a bivariate random variable $(X, \widehat{X})$. Lets consider the sample of the differences $D = (d_1, d_2, ..., d_N)$, where each element $d_i$ is defined as presented in (3.37).

$$d_i = x_i - \widehat{x}_i \tag{3.37}$$

The pairs with $d_i = 0$ are excluded from the test. Accordingly, the number of pairs to perform the test is actualized to $N^{'}$, with $N^{'} \leq N$. Ranks from 1 to $N^{'}$ are defined for each pair $d_i$ following a consecutive and crescent order, using the absolute value $|d_i|$ as criteria. Accordingly, the rank 1 is allocated to the pair with smallest $|d_i|$, the rank 2 is allocated to the pair with the second smallest value of $|d_i|$, and so on. In case of ties, the average of the rank is considered for the elements tied. Once all ranks are assigned, the set of ranks $R = (r_1, r_2, ..., r_N)$ is finally defined using the rule (3.38).

$$\begin{aligned} r_i &= \text{the rank assigned to } (x_i, \widehat{x}_i) \text{ if } d_i > 0 \\ r_i &= \text{the negative of the rank assigned to } (x_i, \widehat{x}_i) \text{ if } d_i < 0 \end{aligned} \tag{3.38}$$

The assumptions underlying the Wilcoxon Signed Rank Test are: (i) each $d_i$ follows a symmetric distribution, (ii) all elements in $D$ are mutually independent and have the same median, (iii) the elements of $D$ are measured in an interval scale. The test may be formulated with one-tailed or two-tailed form. Here the two-tailed form is presented in (3.39).

$$\begin{aligned} H_0 &: E(X) = E(\widehat{X}) \\ H_1 &: E(X) \neq E(\widehat{X}) \end{aligned} \tag{3.39}$$

If the ranking process included ties, the test statistic $TS$ to use in this test is presented in (3.40).

$$TS = \frac{\sum_{i=1}^{N'} r_i}{\sqrt{\sum_{i=1}^{N'} r_i^2}} \tag{3.40}$$

In case of no ties, the test statistic $TS^+$ should be adopted, as defined in (3.41).

$$TS^+ = \sum_{i=1}^{N} (r_i \text{ where } d_i > 0) \tag{3.41}$$

The critical region is defined as presented in (3.42).

$$CR = [0, c_1[\cup]c_2, N'(N'+1)/2] \tag{3.42}$$

The values $c_1$ and $c_2$ are found in the tables with the quantiles for the Wilcoxon Signed Ranks Test Statistic (see [11], pp. 460) to a specified significance level $\alpha$. If $TS$ or $TS^+ \in CR$, the $H_0$ is rejected. If $TS$ or $TS^+ \notin CR$, then the $H_0$ is accepted and there is statistical evidence to assume that both data sets have identical means with a confidence coefficient of $1 - \alpha$.

## 3.6 Distance-based measures

The validation of simulation models can also be performed with the measurement of the distance between the two data sets under assessment. Several distance-based measures exist in literature, not all can be classified as metrics, and for that reason the term "measures" is preferred.

In this section, each data set is used as a vector, as presented in equation (3.43), and all algorithms reviewed aim to found the distance between the two vectors $\mathbf{x}$ and $\widehat{\mathbf{x}}$.

$$\begin{aligned} \mathbf{x} &= (x_1, x_2, x_3, \ldots, x_N)^T \\ \widehat{\mathbf{x}} &= (\widehat{x}_1, \widehat{x}_2, \widehat{x}_3, \ldots, \widehat{x}_N)^T \end{aligned} \tag{3.43}$$

### 3.6.1 Minkowski distance - $d_r$

Minkowski distance is defined in (3.44).

$$d_r(X, \widehat{X}) = \left( \sum_{i=1}^{N} |x_i - \widehat{x}_i|^r \right)^{1/r} \tag{3.44}$$

Minkowski distance has many variants, depending on the value used for the parameter $r$, the two most known variants are the Euclidean distance, and the Manhattan distance. Minkowski distance and its variants have been widely used to assess similarity of images [26, 27].

#### Euclidean Distance - $d_{r=2}$

Euclidean Distance is a particular case of Minkowski distance, when the parameter $r$ assumes the value 2.

$$d_{r=2}(\mathbf{x}, \widehat{\mathbf{x}}) = \left( \sum_{i=1}^{N} |x_i - \widehat{x}_i|^2 \right)^{1/2} \tag{3.45}$$

#### Manhattan Distance - $d_{r=1}$

Manhattan Distance is a particular case of Minkowski distance, when the parameter $r$ assumes the value 1.

$$d_{r=1}(\mathbf{x}, \widehat{\mathbf{x}}) = \sum_{i=1}^{N} |x_i - \widehat{x}_i| \tag{3.46}$$

### 3.6.2 Short Time Series Distance - $d_{STS}$

The Short Time Series Distance was firstly proposed by [34] as a measure of similarity between time series with small number of elements. This measure emerges under the context of comparing DNA microarray data. The $d_{STS}$ is further applied to measure similarity over time series by [29], under the study of clustering of time series data. The mathematic definition of $d_{STS}$ is presented in equation (3.47).

$$d_{STS}(\mathbf{x}, \widehat{\mathbf{x}}) = \sqrt{\sum_{i=1}^{N-1} \left( \frac{\widehat{x}_{i+1} - \widehat{x}_i}{t_{i+1} - t_i} - \frac{x_{i+1} - x_i}{t_{i+1} - t_i} \right)^2} \tag{3.47}$$

As explained by [34], this measure "corresponds to the square root of the sum of the squared differences of the slopes obtained by considering timeseries as linear functions between measurements".

### 3.6.3 Dynamic Time Warping - $DTW$

Dynamic Time Warping $DTW$ is an algorithm which allows the measurement of discrepancy between two data sets. This algorithm was developed under the context of speech recognition, but it may be applied in other research topics, such as the validation of simulation models, as it is suggested in [43, 44].

$DTW$ is a powerful algorithm to identify whether two data sets "match" with each other. As explained in [9], $DTW$ aligns peaks and valleys as much as possible by expanding and compressing the time axis accordingly. This is made by finding the smallest path of distances between two data sets. DTW is applicable for two data sets with different lengths, but as this work is focused on paired data sets, DTW is here considered for data sets with equal length.

Let $\mathbf{A}$ be the $N$ x $N$ matrix where the ($i^{th}$, $j^{th}$) element of matrix $\mathbf{A}$ contains the distance between the two points $d(x_i, \widehat{x}_j)$. The matrix $\mathbf{A}$ is called cost matrix and is defined in (3.48). The distance used to calculate each element of $\mathbf{A}$ is called cost function, and different cost functions may be applied. Examples of possible cost functions are the euclidean distance $d(x_i, \widehat{x}_j) = d_{r=2}(x_i, \widehat{x}_j)$ or the squared error $d(x_i, \widehat{x}_j) = (x_i - \widehat{x}_j)^2$ [43]. In this work, the euclidean distance (see 3.6.1) is always applied as cost function.

$$A = \begin{bmatrix} d(x_1, \widehat{x}_1) & \cdots & d(x_1, \widehat{x}_N) \\ \vdots & \ddots & \vdots \\ d(x_N, \widehat{x}_1) & \cdots & d(x_N, \widehat{x}_N) \end{bmatrix} \tag{3.48}$$

Matrix $\mathbf{A}$ should include values for steps that are considered valid, and not valid steps should be left in blank.

A new matrix $\mathbf{B}$ is then constructed, using the cost matrix $\mathbf{A}$. The ($i^{th}$, $j^{th}$) element of matrix $\mathbf{B}$ stores the minimum cumulative cost to achieve the corresponding position considering the starting point ($i = 1$, $j = 1$). The reasoning applied to find the minimum cumulative cost over each combination of positions is the same as the classic algorithms of dynamic programming for the Shortest Path Problem, such as the Dijkstra algorithm. One way to calculate matrix $\mathbf{B}$ is defined in (3.49), as suggested in [43].

$$\begin{cases} B_{11} = A_{11} \\ B_{ij} = A_{ij} + min(B_{i-1,j-1}, B_{i-1,j}, B_{i,j-1}), \text{ if } i \wedge j \neq 1 \end{cases} \tag{3.49}$$

Once the Matrix $\mathbf{B}$ is defined, the construction of the warping path $\mathbf{W}$ is started. A warping path is a set of $K$ bivariate elements, being $K \in [N, 2N-1]$. Each element $w_k = [i_k^w, j_k^w]$ (with $i_k^w, j_k^w \in [1, N]$) stores the location positions of the elements of matrix $\mathbf{B}$ corresponding to the shortest cumulative path between the first ($i = 1$, $j = 1$) and last ($i = N$, $j = N$) positions. The warping path is thus defined as $W = \langle w_1, w_2, \cdots, w_K \rangle$, subject to the following conditions:

- **Boundary conditions**: $w_1 = [1, 1]$ and $w_K = [N, N]$. This condition obligates the algorithm to start in the first pair of homologous elements, and to finish in the last pair.

- **Continuity**: This condition ensures that all cells chosen from $A$ are adjacent. The condition is formulated considering $w_{k-1} = [i_{k-1}^w, j_{k-1}^w]$ and $w_k = [i_k^w, j_k^w]$, then $i_k^w - i_{k-1}^w \leq 1$ and $j_k^w - j_{k-1}^w \leq 1$.

- **Monotonicity**: The last condition ensures the algorithm to evolve over the matrix $A$. Considering $w_{k-1} = [i_{k-1}^w, j_{k-1}^w]$ and $w_k = [i_k^w, j_k^w]$, then $(i_k^w - i_{k-1}^w > 0 \wedge j_k^w - j_{k-1}^w \geq 0) \vee (i_k^w - i_{k-1}^w \geq 0 \wedge j_k^w - j_{k-1}^w > 0)$.

The distance $DTW$ is given by the square root of the element $B_{NN}$, as defined in

(3.50). Note that the $DTW$ distance does not satisfy the triangle inequality: ($DTW(\mathbf{a},\mathbf{b})+$ $DTW(\mathbf{b},\mathbf{c})$ is not always $\geq DTW(\mathbf{a},\mathbf{c})$). Therefore, $DTW$ is not a metric.

$$DTW(\mathbf{x},\widehat{\mathbf{x}}) = \sqrt{B_{NN}} \tag{3.50}$$

*Note:* it is possible to find in literature ambiguities on the meaning of the information stored within the warping path elements $w_k$. For example, in the study [10], the elements $w_k$ are firstly defined to store the positions $w_k = [i_k^w, j_k^w]$ (as here defined), and latter they are directly applied to calculate the $DTW$ distance, instead of apply the cumulative cost stored in the matrix $\mathbf{B}$ corresponding to the position stored in $w_K$.

As referenced in [10], when the two data series under assessment have different number of elements, the value obtained in (3.50) should be divided by $K$ (the number of elements found in the warping path), to compensate the lengths divergence. In this work, the data series are considered to be of equal lengths. For that reason, the $DTW$ is defined without considering this compensation, as suggested in the study [43] (pp.5).

Lowest values of $DTW$ refer to a better fit between the two data sets. $DTW$ is a scale dependent distance, which does not enable the comparison of this measure over different case studies. Moreover, $DTW$ may return very high values, as they refer to the shortest cumulated distance from $A_{11}$ to $A_{NN}$, the values returned by $DTW$ increase with the number of elements within the data sets to be analyzed.

When $DTW$ is calculated with the cost function $d(x_i,\widehat{x}_j) = (x_i - \widehat{x}_j)^2$ for each element of matrix $\mathbf{A}$, the value obtained is comparable with the euclidean distance $d_{r=2}$ defined in section 3.6.1. A direct comparison with the Manhattan Distance $d_{r=1}$ would never be possible even if the cost function $d(x_i,\widehat{x}_j) = |(x_i - \widehat{x}_j)|$ was adopted to each element of matrix $\mathbf{A}$, as the $DTW$ includes the square root on the cumulative cost achieved in the element $B_{NN}$, as it was presented in equation (3.50).

Therefore, the $DTW$ can only be compared with the euclidean distance distance $d_{r=2}$ when: the two data sets have the same length, the calculus is made exactly as defined in equation (3.50) (i.e. without dividing $\sqrt{B_{NN}}$ by $K$), and the cost function adopted to construct each element of matrix $\mathbf{A}$ is the squared error $d(x_i,\widehat{x}_j) = (x_i - \widehat{x}_j)^2$. The interpretation of this comparison is explained by [10]: (i) when $DTW = d_{r=2}$ it means that the shortest path found in matrix $\mathbf{B}$ relates to the diagonal elements, and so there is no evidence that the estimate data set is lagged from the original data set; (ii) when $DTW < d_2$, it means that the warping path $\mathbf{W}$ has elements outside the diagonal of matrix $\mathbf{B}$, and there is evidence of pattern dissimilarity between the two data sets. In the

second case, the visual interpretation of matrix $\mathbf{B}$ may help the identification of calibration improvements in the model.

## 3.7 Combined measures

Lets consider the following three quantities in (3.51).

$$\begin{aligned} \lambda_{XX} &= \tfrac{1}{N}\sum_{i=1}^{N} x_i^2 \\ \lambda_{\widehat{X}\widehat{X}} &= \tfrac{1}{N}\sum_{i=1}^{N} \widehat{x}_i^2 \\ \lambda_{X\widehat{X}} &= \tfrac{1}{N}\sum_{i=1}^{N} \widehat{x}_i \cdot x_i \end{aligned} \tag{3.51}$$

The acceptable values for the metrics presented in this section are not established in literature yet. In [45] the following reference values are indicated: $0$ in any of the components is the perfect fit between two data series, values below $20\%$ are really good, values between $20\%$ and $30\%$ are considered fair and values above the $30\%$ are cosidered poor.

The measures further detailed are normally used to analyze wave form series, from where the component *Phase* (sometimes called time of arrival) is so important [45, 43]. The meaning of the Phase component is questionable for non wave form series.

### 3.7.1 Sprague & Gear error- $C_{S\&G}$

Sprague & Gear measure considers errors due to *magnitude* and *phase* differences. The error in magnitude $M_{S\&G}$ and the error in phase $P_{S\&G}$ are firstly calculated. These two components are then used to calculate $C_{S\&G}$ using the square root of the sum of the square of the two components. All formulations are detailed in (3.52).

$$
\begin{aligned}
M_{S\&G} &= \sqrt{\frac{\lambda_{XX}}{\lambda_{\widehat{X}\widehat{X}}}} - 1 \\
P_{S\&G} &= \frac{1}{\pi} cos^{-1}\left(\frac{\lambda_{X\widehat{X}}}{\sqrt{\lambda_{XX}\cdot\lambda_{\widehat{X}\widehat{X}}}}\right) \\
C_{S\&G} &= \sqrt{M_{S\&G}^2 + P_{S\&G}^2}
\end{aligned}
\tag{3.52}
$$

The application of this measure can be found in the studies [45, 43, 44]. The values of $C_{S\&G}$ may vary between the range $[0, 1]$, and return asymmetric values (meaning that $C_{S\&G}(X, \widehat{X}) \neq C_{S\&G}(\widehat{X}, X)$). Lower values of $C_{S\&G}$ indicate a better fit of the estimated data set.

### 3.7.2 Russel's error - $C_R$

Russel's error $C_R$ has two main components of error: *magnitude* and *phase*. The phase component $P_R$ is calculated exactly as $P_{S\&G}$. The magnitude component is $M_R$ is calculated differently, as presented in (3.53). Again, the combination of the two components is made using the square root of the sum of the square of the two components.

$$
\begin{aligned}
M_R &= sign(\lambda_{XX} - \lambda_{\widehat{X}\widehat{X}}) \cdot log_{10}\left(1 + \left|\frac{\lambda_{XX} - \lambda_{\widehat{X}\widehat{X}}}{\sqrt{\lambda_{XX}\cdot\lambda_{\widehat{X}\widehat{X}}}}\right|\right) \\
P_R &= \frac{1}{\pi} cos^{-1}\left(\frac{\lambda_{X\widehat{X}}}{\sqrt{\lambda_{XX}\cdot\lambda_{\widehat{X}\widehat{X}}}}\right) \\
C_R &= \sqrt{M_R^2 + P_R^2}
\end{aligned}
\tag{3.53}
$$

The Russel's error is detailed in the studies [43, 44]. The advantage of $C_R$ over $C_{S\&G}$ error is the overcoming of the asymmetry drawback.

### 3.7.3 Normalized Integral Square error - $C_{NISE}$

Normalized Integral Square error $C_{NISE}$ considers three main aspects: *magnitude*, *phase* and *shape*. The cross-correlation $\rho(n)$, which was defined in section 3.4.3, is used to calculate the estimated lag $n_{lag}$. Once $n_{lag}$ is estimated, a shift of $n_{lag}$ is induced to one

of the data sets, and with that change the quantity $\lambda_{X\widehat{X}}(n_{lag})$ is calculated as defined in (3.51). Next the error measures of magnitude $M_{NISE}$, phase $P_{NISE}$ and shape $S_{NISE}$ are calculated. The combined error $C_{NISE}$ is then calculated as the sum of the three components, as it is detailed in expressions (3.54).

$$\begin{aligned} M_{NISE} &= \rho(n_{lag}) - \frac{2\lambda_{X\widehat{X}}(n_{lag})}{\lambda_{XX}+\lambda_{\widehat{X}\widehat{X}}} \\ P_{NISE} &= \frac{2\lambda_{X\widehat{X}}(n_{lag})-2\lambda_{X\widehat{X}}}{\lambda_{XX}+\lambda_{\widehat{X}\widehat{X}}} \\ S_{NISE} &= 1-\rho(n_{lag}) \\ C_{NISE} &= M_{NISE}+P_{NISE}+S_{NISE} = 1-\frac{2\lambda_{X\widehat{X}}}{\lambda_{XX}+\lambda_{\widehat{X}\widehat{X}}} \end{aligned} \tag{3.54}$$

The application of $NISE$ error can be found in the studies [43, 44]. A better estimation of the model is observed to lower values of $C_{NISE}$. A separated analysis of the components is useful to investigate the error arriving from each component, which may endow the researcher with more information on how to improve the model *calibration*.

# Chapter 4

# Measuring the Warping Path Distortion

This chapter includes the proposal of two new performance criteria: Warping Path Distortion- $WPD$ and the Percentage Warping Path Distortion- $PWPD$. The theoretical definition of $WPD$ and $PWPD$ is firstly provided. $PWPD$ is defined as a percentage of $WPD$. The chapter proceeds with the application of both measures to a practical example.

## 4.1 Warping Path Distortion- $WPD$

The Warping Path Distortion $WPD$ is a new performance criterion proposed in this thesis, based on the Dynamic Time Warping $DTW$ algorithm. $WPD$ returns the average distance between the positions stored within the warping path $\mathbf{W}$ with the corresponding nearest positions of the diagonal of matrix $\mathbf{B}$ (3.49).

When the two data sets follow similar patterns over time, the positions stored within the warping path $\mathbf{W}$ are expected to coincide with the position of the diagonal of matrix $\mathbf{B}$. In this case, the value returned by $WPD$ is zero (the distance between the positions stored in $\mathbf{W}$ and the positions corresponding to the diagonal of matrix $\mathbf{B}$ is null).

When the two data sets have different patterns, $WPD$ is expected to return a level on the distortion between the two patterns. Therefore, $WPD$ may be interpreted as the dissimilarity between the two patterns under analysis.

Next, the mathematical formulation of $WPD$ is presented. The nearest position of the diagonal of matrix $\mathbf{B}$ to an arbitrary position stored within the warping path $w_k = [i_k^w, j_k^w]$ is defined as $p_k = [i_k^p, i_k^p]$. The distance between an arbitrary $w_k$ to the corresponding $p_k$ may be visualized by tracing a line amid these two positions, perpendicularly to direction of matrix $\mathbf{B}$ diagonal, see Figure 4.3 visualization 2.

The calculus of $i_k^p$ is presented in (4.1). Note that all elements in diagonal are sym-

metric, thus the elements integrating $p_k$ are both designated as $i_k^p$. A diagonal position $p_k$ may be integrated with values multiples of $0.5$. The set of nearest diagonal positions relating to an arbitrary warping path $W$ is defined as $P = \langle p_1, p_2, \cdots, p_K \rangle$.

$$i_k^p = \left| \frac{i_k^w - j_k^w}{2} \right| + min(i_k^w, j_k^w) \tag{4.1}$$

The distance between an arbitrary $w_k$ to the corresponding nearest position on diagonal of matrix $\mathbf{B}$, $p_k$, is then calculated using the Manhattan Distance $d_{r=1}$ (see section 3.6.1) as described in (4.2).

$$d_{r=1}(w_k, p_k) = |i_k^w - i_k^p| + |j_k^w - i_k^p| \tag{4.2}$$

The Warping Path Distortion $WPD$ is estimated as the average of the distances $d_{r=1}(w_k, p_k)$, considering $k = (2, ..., K-1)$, as described in equation (4.3). The extremity elements $w_1$ and $w_K$ are excluded. These two elements include always the same positions ($w_1 = [1, 1], w_K = [N, N]$) due to the boundary conditions, as specified in section 3.6.3. Therefore, they do not reflect any information on the dissimilarity between the two data sets under assessment.

$$WPD = \frac{1}{K-2} \sum_{k=2}^{K-1} d_{r=1}(w_k, p_k) \tag{4.3}$$

$WPD$ is a scale independent measure. Therefore, $WPD$ may be compared over different case studies provided the length of the vectors is the same. This restriction occurs as $WPD$ depends on the size of the data sets analyzed $N$.

The values returned by $WPD$ are included within the range $WPD \in [0, \frac{(N-1)^2}{K-2}[$. A lower value of $WPD$ is associated with a better fit of the estimated model. The perfect fit would return a $WPD$ value of zero, and a completely distorted pair of vectors would return the worst value of $WPD$: asymptotically $\frac{(N-1)^2}{K-2}$. An example of "the worst case" is provided in Figure 4.1.

## 4.2 Percentage Warping Path Distortion - $PWPD$

The $WPD$ may be used to calculate the correspondent percentage measure, the Percentage Warping Path Distortion $PWPD$. This measure indicates the percentage distortion of a warping path concerning the diagonal of the respective matrix $\mathbf{B}$. This measure is calculated with the division of $WPD$ by the maximum possible value that $WPD$ may

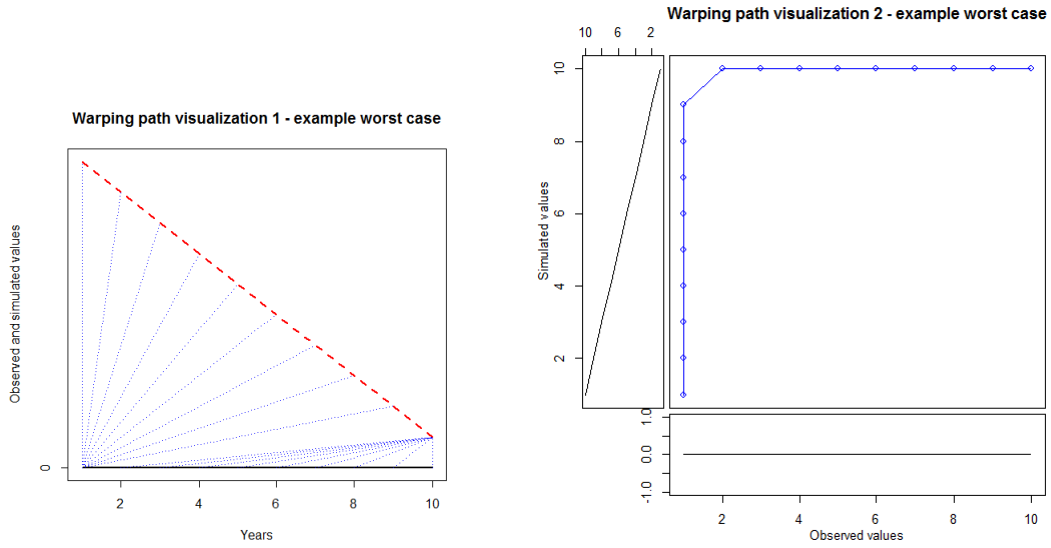


Figure 4.1: Exemplification of a warping path referring to the worst fitting situation.

achieve: $\frac{(N-1)^2}{K-2}$. Accordingly, $PWPD$ is defined as shown in (4.4).

$$PWPD = \frac{WPD}{\frac{(N-1)^2}{K-2}} \tag{4.4}$$

$PWPD$ returns values within the range $[0, 1[$. $PWPD$ does not depend on the number of elements used or the scale of the observations. Therefore, the values obtained with $PWPD$ enable the assessment of single estimations, and are also adequate to perform comparisons over different case studies.

A value of $0$ indicates that the positions integrating the warping path $\mathbf{W}$ are coincident with the positions of the diagonal of matrix $\mathbf{B}$. It also indicates a good similarity between the two patterns under assessment. Therefore, the closest $PWPD$ is to zero, the best the two vectors under assessment fit with each other.

For the worst fitting situation, $PWPD$ would return the asymptotic value of $1$. This value would indicate the maximum distortion of a warping path over the diagonal of matrix $\mathbf{B}$ and the maximum dissimilarity over the two patterns.

## 4.3 A practical example of $WPD$ and $PWPD$

This section presents an exemplification on how two calculate the measure $WPD$. Let us consider the two vectors presented in (4.5), being $\mathbf{x}^T$ a vector containing real observations, and $\widehat{\mathbf{x}}^T$ a vector including a forecast achieved for the observations (these values are used

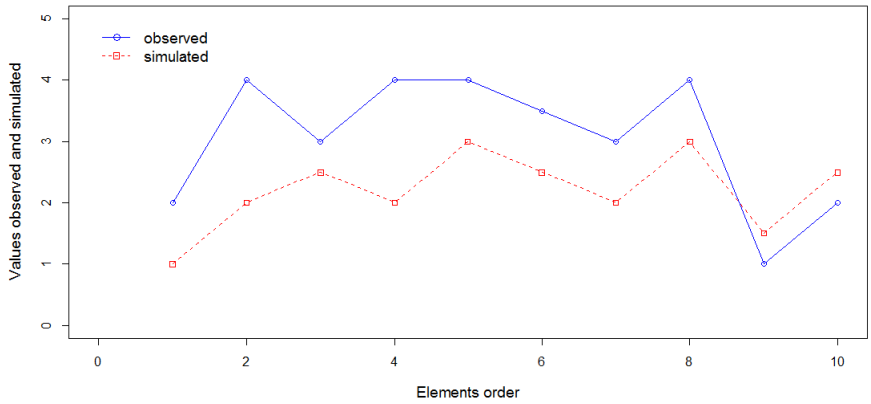


Figure 4.2: Vectors applied under the example.

as an academic example). Figure 4.2 provides the visualization of these vectors.

$$\begin{aligned} \mathbf{x} &= (2.0, 4.0, 3.0, 4.0, 4.0, 3.5, 3.0, 4.0, 1.0, 2.0)^T \\ \widehat{\mathbf{x}} &= (1.0, 2.0, 2.5, 2.0, 3.0, 2.5, 2.0, 3.0, 1.5, 2.5)^T \end{aligned} \tag{4.5}$$

The calculus of the $N$ x $N$ matrix $\mathbf{A}$ is performed using the euclidean distance $d_{r=2}$ as the cost function, considering each combination of two elements from vectors $\mathbf{x}^T$ and $\widehat{\mathbf{x}}^T$. Accordingly, the ($i^{th}$, $j^{th}$) element of matrix $\mathbf{A}$ is calculated as $A_{ij} = d_{r=2}(x_i, \widehat{x}_j) = \sqrt{(x_i - \widehat{x}_j)^2}$. The resulting matrix $\mathbf{A}$ for the example provided is presented in (4.6). This matrix is presented with a repositioning on the positions relative to the estimated vector $\widehat{\mathbf{x}}^T$, in order to make the further visualization of the warping path more intuitive.

$$\mathbf{A} = \begin{array}{c} \\ \widehat{x}_{10} \\ \widehat{x}_9 \\ \widehat{x}_8 \\ \widehat{x}_7 \\ \widehat{x}_6 \\ \widehat{x}_5 \\ \widehat{x}_4 \\ \widehat{x}_3 \\ \widehat{x}_2 \\ \widehat{x}_1 \end{array} \begin{array}{c} \begin{array}{cccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} \end{array} \\ \left( \begin{array}{cccccccccc} 0.5 & 1.5 & 0.5 & 1.5 & 1.5 & 1.0 & 0.5 & 1.5 & 1.5 & 0.5 \\ 0.5 & 2.5 & 1.5 & 2.5 & 2.5 & 2.0 & 1.5 & 2.5 & 0.5 & 0.5 \\ 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 0.5 & 0.0 & 1.0 & 2.0 & 1.0 \\ 0.0 & 2.0 & 1.0 & 2.0 & 2.0 & 1.5 & 1.0 & 2.0 & 1.0 & 0.0 \\ 0.5 & 1.5 & 0.5 & 1.5 & 1.5 & 1.0 & 0.5 & 1.5 & 1.5 & 0.5 \\ 1.0 & 1.0 & 0.0 & 1.0 & 1.0 & 0.5 & 0.0 & 1.0 & 2.0 & 1.0 \\ 0.0 & 2.0 & 1.0 & 2.0 & 2.0 & 1.5 & 1.0 & 2.0 & 1.0 & 0.0 \\ 0.5 & 1.5 & 0.5 & 1.5 & 1.5 & 1.0 & 0.5 & 1.5 & 1.5 & 0.5 \\ 0.0 & 2.0 & 1.0 & 2.0 & 2.0 & 1.5 & 1.0 & 2.0 & 1.0 & 0.0 \\ 1.0 & 3.0 & 2.0 & 3.0 & 3.0 & 2.5 & 2.0 & 3.0 & 0.0 & 1.0 \end{array} \right) \end{array} \tag{4.6}$$

The next step is the calculus of matrix $\mathbf{B}$, as previously defined in expression (3.49). Accordingly, the resulting matrix $\mathbf{B}$ for this case is presented in (4.7).

$$\mathbf{B} = \begin{array}{c} \\ \widehat{x}_{10} \\ \widehat{x}_{9} \\ \widehat{x}_{8} \\ \widehat{x}_{7} \\ \widehat{x}_{6} \\ \widehat{x}_{5} \\ \widehat{x}_{4} \\ \widehat{x}_{3} \\ \widehat{x}_{2} \\ \widehat{x}_{1} \end{array} \begin{array}{c} \begin{array}{cccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} \end{array} \\ \left( \begin{array}{cccccccccc} 5.0 & 6.5 & 7.0 & 8.5 & 10.0 & 10.5 & 9.5 & 11.0 & 11.0 & 10.5 \\ 4.5 & 7.0 & 6.5 & 8.5 & 9.5 & 9.5 & 9.0 & 11.0 & 9.5 & 10.0 \\ 4.0 & 5.0 & 5.0 & 6.0 & 7.0 & 7.5 & 7.5 & 8.5 & 10.5 & 10.5 \\ 3.0 & 5.0 & 5.0 & 7.0 & 9.0 & 8.5 & 7.5 & 9.5 & 10.0 & 9.5 \\ 3.0 & 4.5 & 4.0 & 5.5 & 7.0 & 7.0 & 6.5 & 8.0 & 9.5 & 10.0 \\ 2.5 & 3.5 & 3.5 & 4.5 & 5.5 & 6.0 & 6.0 & 7.0 & 9.0 & 10.0 \\ 1.5 & 3.5 & 4.5 & 6.5 & 8.5 & 9.0 & 9.0 & 11.0 & 11.5 & 11.0 \\ 1.5 & 3.0 & 3.5 & 5.0 & 6.5 & 7.5 & 8.0 & 9.5 & 11.0 & 11.5 \\ 1.0 & 3.0 & 4.0 & 6.0 & 8.0 & 9.5 & 10.5 & 12.5 & 13.5 & 13.5 \\ 1.0 & 4.0 & 6.0 & 9.0 & 12.0 & 14.5 & 16.5 & 19.5 & 19.5 & 20.5 \end{array} \right) \end{array} \tag{4.7}$$

The choice of the warping path $\mathbf{W}$ is then made starting in the element $B_{NN}$ and choosing the nearest small value from $B_{NN}$ backwards, which is highlighted in light gray in (4.7). The "nearest small value backwards" is mathematically defined as $min(B_{i-1,j-1}, B_{i-1,j}, B_{i,j-1})$.

Accordingly, the first element to be included within the warping path $\mathbf{W}$ relates to the position of the matrix $\mathbf{B}$ element $B_{10,10} = 10.5$. This way, $w_K = [10, 10]$. The "nearest small value backwards" from $B_{10,10}$ is to be chosen over the elements $B_{9,9} = 9.5$, $B_{9,10} = 10.0$ and $B_{10,9} = 11.0$. From these three values, the smallest one is the element $B_{9,9} = 9.5$. Therefore, the second element to be included within $\mathbf{W}$ relates to the position of the element $B_{9,9}$, from where $w_{K-1} = [9, 9]$.

This reasoning is applied consecutively until the first element $B_{1,1}$ is reached. At that point the last element $w_1 = [1, 1]$ is included within $\mathbf{W}$. The number of elements integrating the warping path, $K$, is only known when the $\mathbf{W}$ is completely defined. The visualization of the warping path found for this example is provided in Figure 4.3.

Lets consider the elements composing the warping path $\mathbf{W}$, which refer to positions of the elements of matrix $\mathbf{B}$. The next step is the calculus of the set of nearest diagonal positions $\mathbf{P}$ corresponding to $\mathbf{W}$. For this example, the warping path $\mathbf{W}$ and the corresponding $\mathbf{P}$ are presented in (4.8).

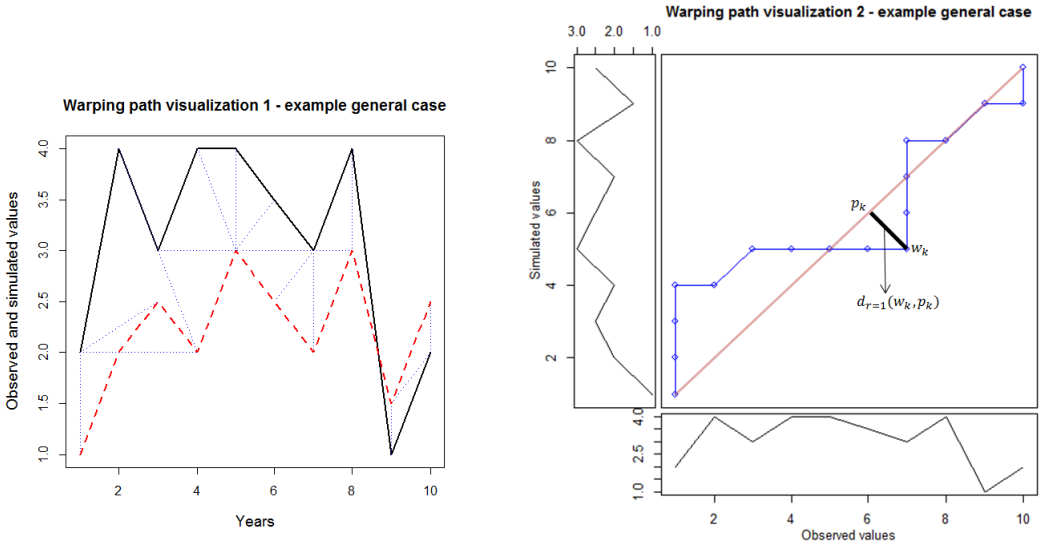


Figure 4.3: Visualization of the warping path for the example *general case*.

$$
\begin{cases}
w_1 = [1,1] \Rightarrow p_1 = [1,1] \\
w_2 = [1,2] \Rightarrow p_2 = [1.5,1.5] \\
w_3 = [1,3] \Rightarrow p_3 = [2,2] \\
w_4 = [1,4] \Rightarrow p_4 = [2.5,2.5] \\
w_5 = [2,4] \Rightarrow p_5 = [3,3] \\
w_6 = [3,5] \Rightarrow p_6 = [4,4] \\
w_7 = [4,5] \Rightarrow p_7 = [4.5,4.5] \\
w_8 = [5,5] \Rightarrow p_8 = [5,5] \\
w_9 = [6,5] \Rightarrow p_9 = [5.5,5.5]
\end{cases}
\quad
\begin{cases}
w_{10} = [7,5] \Rightarrow p_{10} = [6,6] \\
w_{11} = [7,6] \Rightarrow p_{11} = [6.5,6.5] \\
w_{12} = [7,7] \Rightarrow p_{12} = [7,7] \\
w_{13} = [7,8] \Rightarrow p_{13} = [7.5,7.5] \\
w_{14} = [8,8] \Rightarrow p_{14} = [8,8] \\
w_{15} = [9,9] \Rightarrow p_{15} = [9,9] \\
w_{16} = [10,9] \Rightarrow p_{16} = [9.5,9.5] \\
w_{17} = [10,10] \Rightarrow p_{17} = [10,10]
\end{cases}
\tag{4.8}
$$

The $WPD$ value obtained for this example is $1.133$. This value is primarily a measure of the distortion incurred by the warping path concerning the diagonal of matrix $\mathbf{B}$.

$WPD$ can also be interpreted as a measure of pattern dissimilarity between the estimated and the observed data sets. For this example, the values obtained with $WPD$ are constrained by the interval $WPD \in [0, \frac{(10-1)^2}{17-2}[ \Rightarrow [0, 5.4[$. The value obtained can be considered low, indicating a small dissimilarity of the two patterns analyzed.

Once the value $WPD$ is estimated, the calculus of the measure $PWPD$ is obtained dividing $WPD$ by $5.4$, resulting in the value $PWPD = 0.210$. Now, the dissimilarity is easily interpreted, the two patterns are divergent in approximately $21\%$, that is not a low value at all. Therefore, if the estimated obtained with $WPD$ is passible to originate some ambiguities on the assessment of a

single forecast, the $PWPD$ is clear, as it is interpreted similarly as a percentage measure varying between $0\%$ and assyntotically $100\%$.

# Chapter 5

# Practical Exploration of Performance Criteria

This chapter applies the performance criteria reviewed in chapters 3 and 4 to five data sets. The data analyzed refer to the results obtained with a simulation model developed by the author to study demographic aspects of Portuguese firms, here called *PoFi* (**Po**rtuguese **Fi**rms). The five experiments ascribe to the number of firms on Portuguese geographic areas *Norte*, *Centro*, *Lisboa*, *Alentejo* and *Algarve*, between the years 1985 and 2009.

*PoFi* was developed using cellular automata techniques, with main inspiration on the Conway's Game of Life [15]. In the light of this work, the model is useful to apply the resulting data over the reviewed performance criteria. Thus, the description on the algorithms used to construct the model is provided just summarily. An important note is that some experiments of *PoFi* further analyzed intensionally relate to bad forecasts, as the main goal of this section is to see how the performance criteria behave under different situations.

The access to historic data used in this case study was gently provided by "*GEP do Ministério do Trabalho e Solidariedade Social, Portugal*", and the author assumes total responsibility for the interpretation made under this study.

## 5.1 *PoFi* Construction

*PoFi* is based on a four dimensional array. The first and second dimensions represent the geographic location of firms. These dimensions may be interpreted as coordinates of the firms' position, with reference to the simplified Portugal map defined as a rectangular matrix, here called *base matrix*. The number of positions in *base matrix* may be changed automatically, a higher number of positions is associated with more accurate results and higher computational effort. Each firms' position in *base matrix* is linked with the correspondent Nomenclature of Territorial Units for Statistics - NUTS. *PoFi* includes the following level $II$ Portuguese NUTS: *Norte*, *Centro*, *Lisboa*,

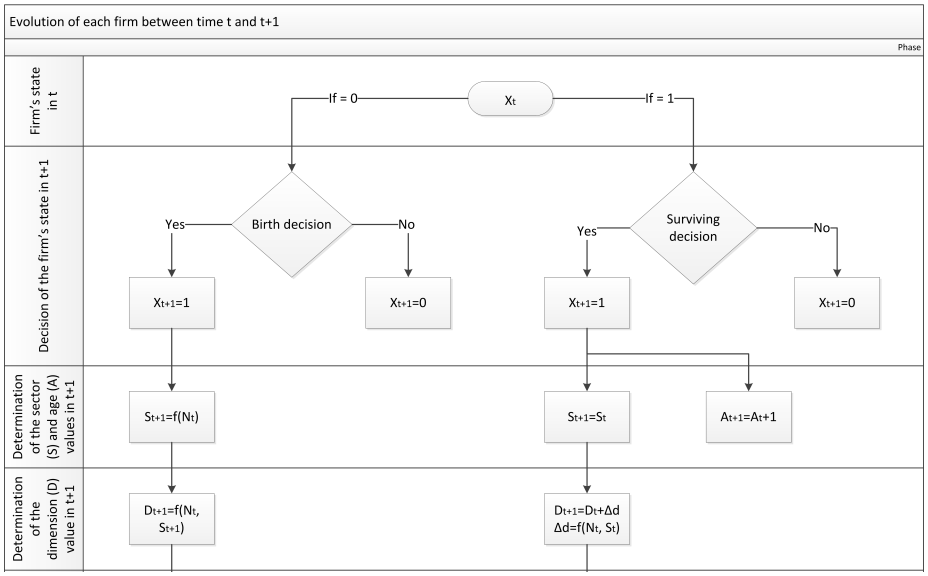


Figure 5.1: Diagram representing the algorithm followed by *PoFi*. This diagram explores the transition of one firm during one iteration (between t and t+1). State of the firm: X, sector of the firm: S, age: A, geographical region: N, dimension: D, variation of dimension: $\Delta$d.

*Alentejo* and *Algarve*.

The third dimension refers to the time evolution in years. The fourth dimension includes information on the following categories: (i) state of firms - alive or dead, (ii) age in years, (iii) sector including a) agriculture and fisheries, b) industry, c) construction, d) services, (iv) dimension that encompasses four levels of firms' size.

The model evolves according to micro rules established for each cell on the four dimensional array, following an annual periodicity. The Conway's Game of life [15] was used as a starting point of *PoFi*, from where more variables and rules were considered. The rules specified for each cell within an arbitrary iteration (i.e. between time $t$ and time $t + 1$) are described in Figure 5.1, which is organized in four main phases.

The *first phase* identifies the current state of a firm. When the firm is alive its state is 1 ($X_t = 1$), and when the firm is dead its state is 0 ($X_t = 0$). Depending on the current state, the model follows different paths.

### 5.1.1 Algorithm used when a firm is dead

Considering that the current state is dead, the *second phase* will decide the potential success of the firm to be born or to continue dead, considering the number of living firms close to the firm in study. The close firms considered are those within a radius of one cell. If the firm's state is kept in 0, null values are kept for all characteristics. In case the firm's state is changed to 1, the algorithm evolves to the next phase.

The *third phase* selects the activity sector for the born firm. The model includes different probabilities for the selection of each sector of activity considering the NUTS of the firm. These probabilities are the average proportion of observed firms of each sector within each zone considering the aggregation of all years. In this phase, a uniform random number is generated, and the selection of the sector is carried out with the comparison of the random number obtained and the probability values found.

The *fourth phase* considers two features to define the probability associated to the dimension definition: NUTS and the sector of activity (i.e. it considers the average proportion of observed firms of each dimension within each sector within each zone, in real data). This phase also uses uniform random numbers to perform the decision of the dimension of the new firm.

At this stage, all characteristics of a new firm are defined. The respective information is saved and used on a new iteration.

### 5.1.2 Algorithm used when a firm is alive

When the initial state of a firm is $1$, meaning that the firm is alive, the *second phase* decides on the survival of this firm. The survival decision is made considering the number of firms living within the radius of one cell to the cell under assessment. If the firm dies, all characteristics are set to zero. If the firm survives, the next phase is pursued.

The *third phase* defines information on two characteristics: sector and age. The sector of a surviving firm is not changed. The generality of the defined classes is so pervasive that it is considered very unlikely that a firm performs this change. The variable age is incremented in one unit every time a firm survives (note that when a firm is born, its age is still zero, and that a firm only have the first year completed when it achieves the first year of survival).

The *fourth phase* defines the new value of the firm's dimension. This is formulated considering the assumption that a firm will have a maximum absolute size variation of one per iteration (i.e. in each iteration the dimension can be incremented by $-1$, $0$ or $1$). This variation is determined with base on the proportion of observed firms changing its dimension as a probability.

## 5.2 Results

The model provided results concerning each of the characteristics simulated. Next, the results obtained concerning the number of living firms in each zone are analyzed. Figure 5.2 presents the number of living firms simulated alongside with the respective historic records.

The next sections include the validation assessment for each zone separately. The Performance criteria further analyzed were calculated considering values in $10^3$ units, using the R software for statistics [40]. The Parametric tests should not be calculated for this example, as the samples have a size of $N = 25$, and the condition of normality is not ensured. Nevertheless, and knowing that

**Error Based Measures**

| | |
|---|---|
| $ME$ | $-9.199$ |
| $MAE$ | 26.863 |
| $MSE$ | 911.899 |
| $RMSE$ | 30.198 |
| $U_1$ | 0.174 |
| $U_2$ | 0.356 |
| $MPE$ | $-0.283$ |
| $MAPE$ | 0.433 |
| $MASE$ | 6.568 |

**Information Theory Measures**

| | | |
|---|---|---|
| $H_S$ | $H_S(X)$ | 3.138 |
| | $H_S(X, \widehat{X})$ | 3.865 |
| $I(X, \widehat{X})$ | | 2.478 |
| $NMI(X, \widehat{X})$ | | 0.789 |
| $Id(X, \widehat{X})$ | | 0.210 |

**Distance Based measures**

| | |
|---|---|
| Euclidean distance | 150.988 |
| Manhattan distance | 671.573 |
| $\mathrm{d}_{STS}$ | 64.476 |
| DTW | 21.045 |

**Distortion Path Measures**

| | |
|---|---|
| DWP | 5.432 |
| PDWP | 0.415 |

**Parametric Tests**

| | | |
|---|---|---|
| Pearson | $r$ | 0.396 |
| | p-value | 0.049 |
| $R^2$ | | 0.157 |
| $\rho(n)$ | $\rho(n_{lag})$ | 0.415 |
| | $n_{lag}$ | 2 |

**Nonparametric Tests**

| | | |
|---|---|---|
| Sign test | TS | 14 |
| | $CR_{\alpha=0.05}$ | $[0, 7.6[\cup]17.4, 25]$ |
| Spearman | $r_S$ | 0.397 |
| | $CR_{\alpha=0.05}$ | $]0.337, +\infty[$ |
| Kendall | $\tau$ | 0.29 |
| | p-value | 0.049 |
| Wilcoxon | TS | 105 |
| | p-value | 0.127 |

**Combined measures**

| | | |
|---|---|---|
| S&G | $M_{S\&G}$ | $-0.048$ |
| | $P_{S\&G}$ | 0.318 |
| | $C_{S\&G}$ | 0.322 |
| Russel | $M_R$ | $-0.041$ |
| | $P_R$ | 0.318 |
| | $C_R$ | 0.320 |
| NISE | $M_{NISE}$ | $-0.525$ |
| | $P_{NISE}$ | 0.001 |
| | $S_{NISE}$ | 0.585 |
| | $C_{NISE}$ | 0.060 |

Table 5.1: Performance Criteria summary for the number of firms in *Norte*.

this restriction is relaxed, the values on these criteria are presented and discussed.

Only two normalized information measures are further presented, as they all lead to similar conclusions and the analysis of one $NMI$ and one $I_d$ is considered sufficient to demonstrate how these measures behave. Following the same reasoning on the calculus of Percentage Error $PE$ (always in order to the observed values), the $NMI$ further presented refer to the division of the mutual information $MI(X, \widehat{X})$ by the entropy of the observed data set $H(X)$. The second normalized information measure analyzed is the corresponding information distance $I_d = 1 - NMI$.

### 5.2.1 Number of firms in *Norte*

The performance criteria obtained for the number of firms in *Norte* are presented in Table 5.2.1, and the Warping path obtained is illustrated in Figure 5.2.1.

The analysis of the number of firms in *Norte* is made with a single estimated data set. Consequently, isolated values of $ME$, $MAE$, $MSE$ and $RMSE$ do not endow a proper conclusion on the goodness of fit of this feature. These measures are not comparable over the remainder ex-

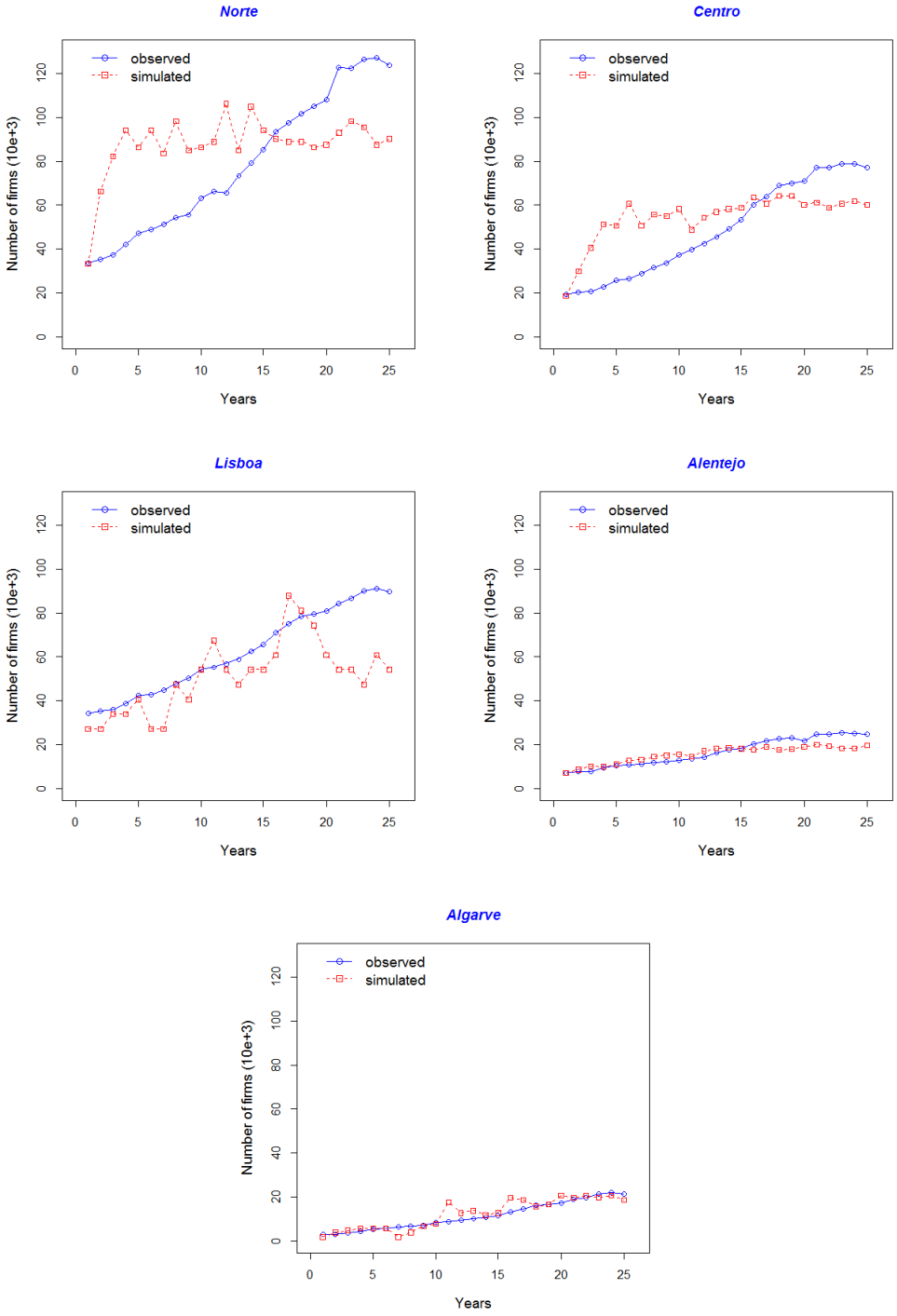


Figure 5.2: Simulated and observed data from the Portuguese case study.

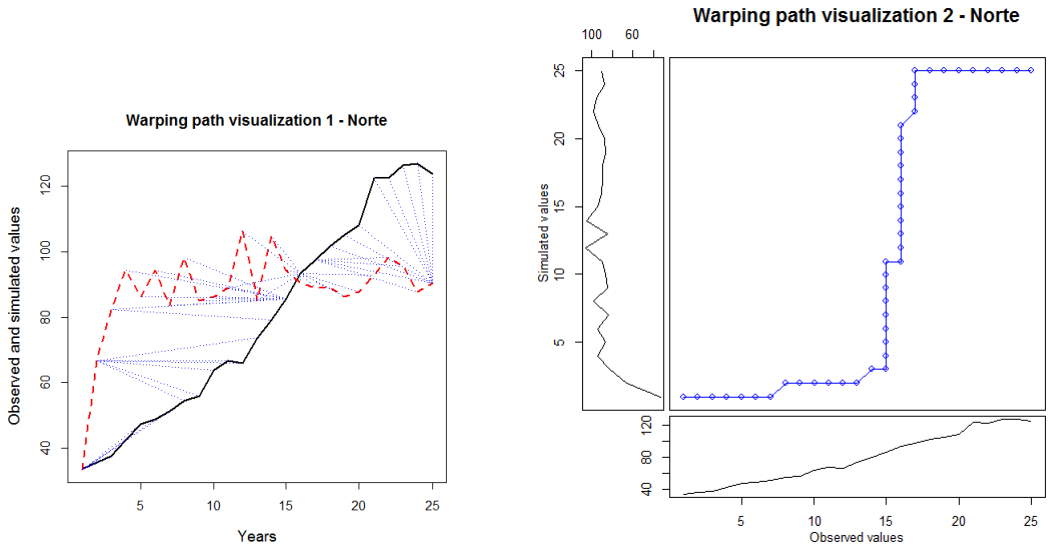


Figure 5.3: Warping path visualization for the number of firms in *Norte*.

periments presented in this section, they would only be comparable for different forecasts of the same feature. The same reasoning is applicable for the remainder experiments on *Centro*, *Lisboa*, *Alentejo* and *Algarve*.

The value obtained with the Theil's coefficient for forecast accuracy $U_1$ returned the worst value for the *Norte* forecast over the five experiments conducted. The value of $0.174$ cannot be considered as a good estimate, as it is not close the reference value of $0$. The coefficient $U_2$ returned a value lower than $1$, which indicates that the estimation obtained has lower standard error as the naive no-change extrapolation. If the value obtained with $U_2$ was larger than $1$, the model should be rejected as in that case the standard error obtained with the forecast would be worst than the simples no-change extrapolation.

The negative sign obtained with $ME$ and $MPE$ indicates that the estimation of *Norte* feature returns in average higher values than the ones observed. This conclusion may be verified with the visualization of Figure 5.2, and is important for future calibration of the model. These measures, alongside with $M_{S\&G}$ and $M_R$ allow to draw similar conclusions.

The percentage-based error measures are more intuitive as they allow the goodness of fit assessment for a single estimation and they are comparable over experiments on different features. The mean percentage error of $-28.3\%$ indicates that $\widehat{X}$ values are in average $28.3\%$ above $X$ (as it is observed till the $15^{th}$ year considered). The mean absolute percentage error indicates that in average the estimations $\widehat{X}$ are distanced from the historic record $X$ in $43.3\%$, which is a high value showing the bad fitness obtained for this feature.

The *Norte* case has the worst MASE value over the five experiments under assessment in this section. As the range of MASE is $[0, +\infty[$, the value $6.568$ here obtained purports more

information when compared with other MASE values. As MASE is scale-free, the values of this measure are comparable over results on different features.

The entropy value obtained for the observed data set is similar to the joint entropy over the two data sets. However, the mutual information shows that the two series are not similar, since the mutual information is considerably lower than the joint entropy. The normalized mutual information allows a more intuitive assessment on the quantity of information shared by the two data sets, that can be interpreted as a percentage of $78.9\%$. This means that the two data set share $78.9\%$ of information, and this value is too low to be considered a good result (the perfect fit would return the value $100.0\%$). As it was expected, the normalized information distance, aproximately $21.0\%$, is too high to be considered a good estimation (the perfect fit would return the value of $0.0\%$. These measures indicate that the simulated data set is a bad representation of reality concerning the quantity of information.

The parametric test on the Pearson's correlation coefficient indicates a positive correlation of $r = 0.396$, with a statistical significance of $0.049$. The coefficient of determination is low, meaning that the simulated data explains only a small portion of the variability observed in reality. The cross correlation shows that the simulated data would better fit the historic records with a lag of two temporal units.

The Sign test does not reject the null hypothesis of equality of $X$ and $\widehat{X}$ medians, thus, considering a statistical significance of $\alpha = 0.05$, these medians are similar.

The Spearman correlation returns a positive and statistically significant correlation, considering $\alpha = 0.05$. This evaluation may concluded as $r_S \in CR_{\alpha=0.05}$, thus the null hypothesis is rejected. Note that for an $\alpha = 0.01$, this correlation is no longer significant. For this case, the Spearman's correlation test should be adopted instead of Pearson, as the size of the data sets is small.

The Kendall's $\tau$ obtained indicates a moderate and positive correlation, with a p-value of $0.049$.

The Wilcoxon test for means does not allow the rejection of the null hypothesis $E(X) = E(\widehat{X})$, as the p-value obtained is higher than $\alpha = 0.05$. Therefore, there is no statistical evidence to assume that the means followed by $X$ and $\widehat{X}$ are different, and the null hypothesis is kept.

The values returned by the distances $d_{r=2}$ and $d_{r=1}$ seem to be high, but without an equivalent comparison, the information these criteria have is limited. High values of these distances indicate that the two data sets are in average separated from each other by high amplitudes. In fact, the visualization of Figure 5.2 for the *Norte* case indicates this same conclusion.

The distance $d_{STS}$ provides a different assessment. It indicates a measurement on the average amplitude divergence over consecutive temporal moments, between the two data sets. In this case, the square root of the sum of the squared differences of the slopes between temporal moments returned the value $d_{STS} = 64,476$. The complex interpretation of this measure is one of its drawbacks.

The values obtained with $DTW$ does not include much information without a comparison of the same measure with a different forecast for the same case study. One of the main drawbacks of $DTW$ is its dependency on the observations scale and on the length of the data sets. Note that the cost function used to construct the matrix $\mathbf{A}$ was not the squared error, and therefore the $DTW$ values here presented are not comparable with the euclidean distance $d_{r=2}$.

Concerning the combined measures, the phase components $P_{S\&G}$ and $P_R$ are suitable to assess lag on series with sinusoidal behavior, or wave form, as explained in [45]. This is not the case for the experiments conducted in this section, and so, these phase indicators are meaningless. Concerning the phase component of NISE, it relates with the results obtained for the number of periods of lag with $\rho(n)$. For the experiment *Norte*, as $n_{lag}$ was set to 2, the $P_{NISE}$ returns a value slightly different from zero. For the remainder experiments, where $n_{lag} = 0$, the $P_{NISE} = 0$ as well.

The magnitude components $M_{S\&G}$ and $M_R$ were analyzed alongside with the sign returned by ME and MPE. These magnitude components return values within the range $]-1, 1[$, and are frequently interpreted as a percentage of magnitude discrepancy on the series under analysis. Concerning the $M_{NISE}$, it returns values within the range $[-1, 0]$, from where it provides information that should be interpreted as the absolute magnitude percentage deviation. For the *Norte* case, the $M_{NISE}$ returns a percentage deviation of $25,5\%$, that is a really bad estimation, and more severe than the MAPE estimation.

The shape component $S_{NISE}$ returns values within $[0, 1]$, being $0$ the best possible result indicating a good pattern similarity and $1$ the worst possible result indicating a bad similarity. For the *Norte* estimation $S_{NISE} = 0.585$, that is a bad pattern similarity. This measure may be compared with $PDWP$, as they assess the same characteristic although using different methodologies. Note that $PDWP$ returns values within the same range and with similar interpretation as $S_{NISE}$. The result obtained with $PDWP$ is however less severe than $S_{NISE}$, and these two measures indicate a considerably bad pattern similarity.

The value returned by $WPD$ was $5,432$ out of a maximum of $13.091$. This result indicates a bad pattern similarity between the two data sets. This dissimilarity is also observable in Figure 5.2.1, showing that the Warping Path is highly divergent from the reference position from diagonal. The value $PDWP$ quantifies this divergence in $41.5\%$.

### 5.2.2 Number of firms in *Centro*

The performance indexes obtained for the number of firms in the zone Centro are presented in Table 5.2.2, and the warping path obtained for the two time series is shown in Figure 5.2.2.

The measures ME, MAE, MSE and RMSE are lower for the estimated number of firms in *Centro* than in *Norte*. Although this comparison is tempting, it should not be made, as scale-based errors are not suitable to assess estimates over different features. The error-based measures

**Error Based Measures**

| | |
|---|---|
| $ME$ | $-5.643$ |
| $MAE$ | 14.749 |
| $MSE$ | 289.460 |
| $RMSE$ | 17.014 |
| $U_1$ | 0.156 |
| $U_2$ | 0.319 |
| $MPE$ | $-0.289$ |
| $MAPE$ | 0.411 |
| $MASE$ | 5.757 |

**Information Theory Measures**

| | | |
|---|---|---|
| $H_S$ | $H_S(X)$ | 3.120 |
| | $H_S(X, \widehat{X})$ | 3.852 |
| $I(X, \widehat{X})$ | | 2.465 |
| $NMI(X, \widehat{X})$ | | 0.790 |
| $Id(X, \widehat{X})$ | | 0.210 |

**Distance Based measures**

| | |
|---|---|
| Euclidean distance | 85.068 |
| Manhattan distance | 368.722 |
| $d_{STS}$ | 28.981 |
| DTW | 14.101 |

**Distortion Path Measures**

| | |
|---|---|
| DWP | 5.163 |
| PDWP | 0.385 |

**Parametric Tests**

| | | |
|---|---|---|
| Pearson | $r$ | 0.688 |
| | p-value | 0.000 |
| $R^2$ | | 0.473 |
| $\rho(n)$ | $\rho(n_{lag})$ | 0.688 |
| | $n_{lag}$ | 0 |

**Nonparametric Tests**

| | | |
|---|---|---|
| Sign test | TS | 15 |
| | $CR_{\alpha=0.05}$ | $[0, 7.6[\cup]17.4, 25]$ |
| Spearman | $r_S$ | 0.791 |
| | $CR_{\alpha=0.05}$ | $]0.337, +\infty[$ |
| Kendall | $\tau$ | 0.628 |
| | p-value | 0.000 |
| Wilcoxon | TS | 99 |
| | p-value | 0.090 |

**Combined measures**

| | | |
|---|---|---|
| S&G | $M_{S\&G}$ | $-0.039$ |
| | $P_{S\&G}$ | 0.318 |
| | $C_{S\&G}$ | 0.321 |
| Russel | $M_R$ | $-0.034$ |
| | $P_R$ | 0.318 |
| | $C_R$ | 0.318 |
| NISE | $M_{NISE}$ | $-0.264$ |
| | $P_{NISE}$ | 0.000 |
| | $S_{NISE}$ | 0.312 |
| | $C_{NISE}$ | 0.049 |

Table 5.2: Performance Criteria summary for the number of firms in *Centro*.

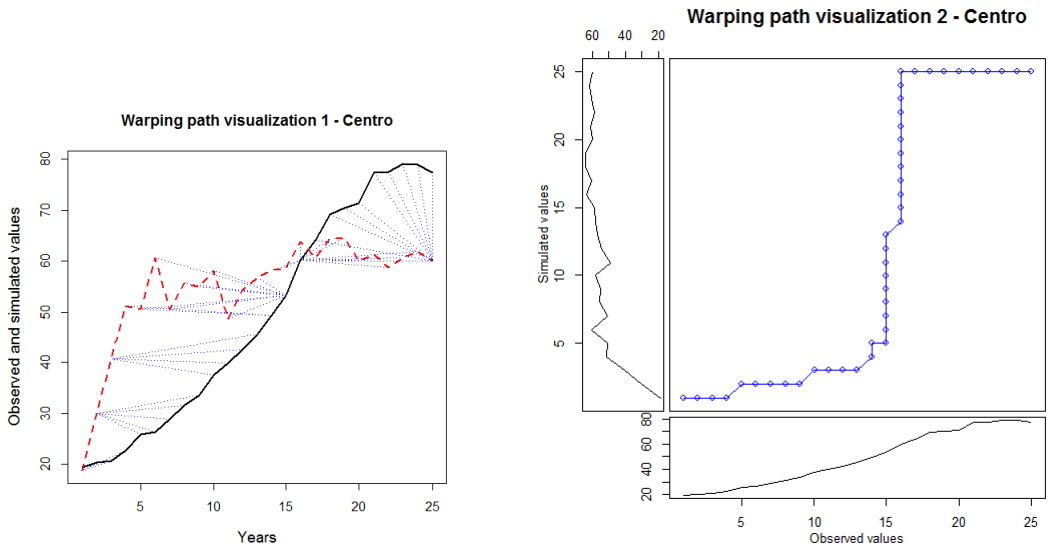


Figure 5.4: Warping path visualization for the number of firms in *Centro*.

suitable to address this comparison are MPE, MAPE and MASE.

The value obtained with $U_1$ is slightly better than the correspondent value obtained for *Norte*, but it cannot be considered a good estimate since it is not close to the reference $0$. The value obtained with $U_2$ is lower than the reference $1$, indicating that this forecast has lower standard error than the simple naive no-change extrapolation.

The negative sign obtained with ME, MPE, $M_{S\&G}$ and $M_R$ indicate that this estimation returns in average higher values than the respective historic records. Figure 5.2.2 shows that the number of estimated firms is always higher than the historic record till the $16^{th}$ year simulated. Future calibration of this model should be conducted to improve this misalignment.

The MPE for *Centro* shows a divergence between estimated and observed data slightly higher than the one observed in *Norte* case, with a result of $-28.9\%$, indicating that $\widehat{X}$ is in average $28.9\%$ above $X$. Therefore, with the MPE criterion, *Centro* would be a worst estimation in comparison with *Norte*.

MAPE shows the contrary conclusion. The MAPE value for *Centro* is better than the MAPE value for *Norte*. This disagreement is perfectly understandable as MPE is a measure that offsets positive and negative values of error. A better comparison over different estimations is achieved with the MAPE measure, which does not offsets positive or negative values.

The conclusion taken with MASE is similar to the one obtained with MAPE, the estimation for *Centro* are better than the ones for *Norte*. This conclusion is taken as the MASE value for *Centro* is lower than the one observed for *Norte*. MAPE has the advantage of being defined within the closed range $[0, 1]$, which is a more intuitive assessment.

The results obtained for Centro concerning the information theory measures are similar to the ones drawn for the Norte experiment. The normalized mutual information is perhaps the most intuitive information measure over the four analyzed, which indicates that only $79.0\%$ of the information contained in $X$ is contained in $\widehat{X}$.

The Pearson's correlation tests indicates that both estimated and historic data sets for Centro are positively correlated, with a high statistical significance (p-value= $0.000$). The coefficient of determination indicates that the variability explained by the model is $0.473$ out of $1$. The cross correlation indicates that the simulated data would not provide a better fit with any lag on temporal units.

The $TS$ obtained with the Sign test is not included within the critical region defined, from where the null hypothesis that tests the equality of $X$ and $\widehat{X}$ medians is not rejected, meaning that the two medians are considered similar for a statistical significance of $\alpha = 0.05$.

The Spearman's correlation returns a positive and statistically significant correlation, for $\alpha = 0.05$. This conclusion is taken as $r_S \in CR_{\alpha=0.05}$, and consequently the null hypothesis testing no correlation is rejected. For an $\alpha = 0.01$ this correlation keeps to be significant ($CR_{\alpha=0.01} = ]0.466, +\infty[$).

The Kendall's $\tau$ obtained indicate a strong and positive correlation, with a p-value of $0.000$,

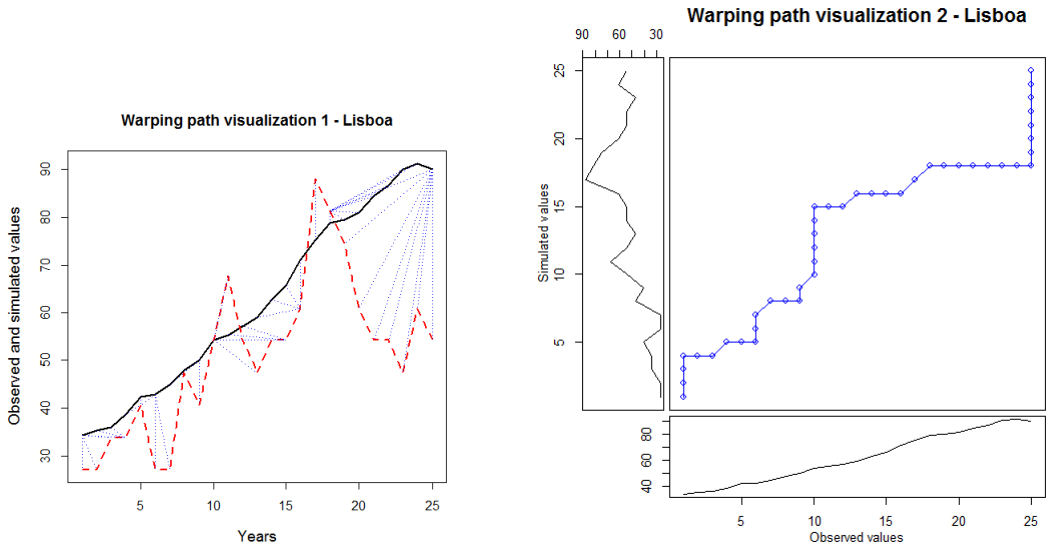


Figure 5.5: Warping path visualization for the number of firms in *Lisboa*.

meaning that the two data series are positively correlated for any significance coefficient $\alpha$ considered.

The Wilcoxon test for means do not reject the null hypothesis $E(X) = E(\widehat{X})$, as the p-value obtained is higher than $\alpha = 0.05$. Therefore, the null hypothesis of equality of the means followed by the two data series is kept.

The distances $d_{r=2}$ and $d_{r=1}$ returned lower values for *Centro* in comparison with *Norte*. This means that the estimated and observed data sets are closer for the *Centro* experiment than for the *Norte* experiment. This result suggests the *Centro* estimations to be better than the *Norte* ones.

The distance $d_{STS}$ indicates that the slope divergence between each pair of consecutive temporal moments in *Centro* is lower than in *Norte*. Therefore, this criterion suggests *Centro* to be a better estimation than *Norte*. The comparison of the $DTW$ values indicate similar conclusions.

As for the *Centro* experiment, $n_{lag} = 0$, the phase component of NISE is 0 as well. The $M_{NISE}$ indicates that the two data series are deviated from one another in $26,4\%$, which indicates high diferences concerning this characteristic.

The shape component $S_{NISE}$ for this case was $0.312$, indicating that the patterns are not similar, although it is a better result than the one obtained for *Norte*. A similar conclusion is obtained with $PDWP$, with a value of $0.385$.

### 5.2.3 Number of firms in *Lisboa*

The performance indexes obtained for the number of firms in *Lisboa* are presented in Table 5.2.3. The Warping path obtained for this experiment is illustrated in Figure 5.2.3.

**Error Based Measures**

| | |
|---|---|
| $ME$ | 11.279 |
| $MAE$ | 13.477 |
| $MSE$ | 320.431 |
| $RMSE$ | 17.901 |
| $U_1$ | 0.151 |
| $U_2$ | 0.275 |
| $MPE$ | 0.166 |
| $MAPE$ | 0.200 |
| $MASE$ | 5.551 |

**Information Theory Measures**

| | | |
|---|---|---|
| $H_S$ | $H_S(X)$ | 3.171 |
| | $H_S(X, \widehat{X})$ | 3.858 |
| $I(X, \widehat{X})$ | | 2.481 |
| $NMI(X, \widehat{X})$ | | 0.782 |
| $Id(X, \widehat{X})$ | | 0.218 |

**Distance Based measures**

| | |
|---|---|
| Euclidean distance | 89.503 |
| Manhattan distance | 336.929 |
| $\mathrm{d}_{STS}$ | 50.180 |
| DTW | 17.642 |

**Distortion Path Measures**

| | |
|---|---|
| DWP | 2.219 |
| PDWP | 0.158 |

**Parametric Tests**

| | | |
|---|---|---|
| Pearson | $r$ | 0.699 |
| | p-value | 0.000 |
| $R^2$ | | 0.488 |
| $\rho(n)$ | $\rho(n_{lag})$ | 0.699 |
| | $n_{lag}$ | 0 |

**Nonparametric Tests**

| | | |
|---|---|---|
| Sign test | TS | 3 |
| | $CR_{\alpha=0.05}$ | $[0, 7.6[\cup]17.4, 25]$ |
| Spearman | $r_S$ | 0.718 |
| | $CR_{\alpha=0.05}$ | $]0.337, +\infty[$ |
| Kendall | $\tau$ | 0.543 |
| | p-value | 0.000 |
| Wilcoxon | TS | 287 |
| | p-value | 0.000 |

**Combined measures**

| | | |
|---|---|---|
| S&G | $M_{S\&G}$ | 0.216 |
| | $P_{S\&G}$ | 0.318 |
| | $C_{S\&G}$ | 0.385 |
| Russel | $M_R$ | 0.144 |
| | $P_R$ | 0.318 |
| | $C_R$ | 0.349 |
| NISE | $M_{NISE}$ | $-0.256$ |
| | $P_{NISE}$ | 0.000 |
| | $S_{NISE}$ | 0.301 |
| | $C_{NISE}$ | 0.045 |

Table 5.3: Performance Criteria summary for the number of firms in *Lisboa*.

$ME$, $MPE$, $M_{S\&G}$ and $M_R$ return positive values, meaning that the estimated data set for *Lisboa* has, in average, lower values than the respective observed data set.

The measure $U_1$ indicates that this forecast is a bad representation of the correspondent historic record, as it is not close to the reference value of $0$. The measure $U_2$ returned a lower value than the reference $1$, indicating that this model should not be reject when compared with the simple naive no-change extrapolation.

All the comparable error-based measures over different features, $U_1$, $U_2$, MPE, MAPE and MASE, show that the simulation model obtained a better fit for the estimated number of firms in *Lisboa* than in *Norte* or *Centro*.

The information theory measures returned similar results for this experiment than for the *Norte* and *Centro* experiments.

The results on Pearson's correlation test indicate that the estimated model is strongly correlated with the real observations, with $r = 0.699$ and p-value of $0.000$. The measure $R^2$ indicates that the variability explained by the model is $0.488$ out of $1$, and the cross correlation did not identify any temporal lag on the estimated data set.

The Sign test $TS$ obtained is not within the critical region defined for a statistical significance of $\alpha = 0.05$, thus the null hypothesis testing the equality of $X$ and $\widehat{X}$ medians is not rejected, and the medians are considered similar.

The Spearman's correlation returns a strong positive correlation, that is statistically significant considering both $\alpha = 0.05$ and $\alpha = 0.01$ ($CR_{\alpha=0.01} = ]0.466, +\infty[$). This conclusion is taken as $r_S \in CR_{\alpha=0.05}$, from where the null hypothesis of $r_S = 0$ is rejected.

For the *Lisboa* estimations obtained, the Kendall's $\tau$ returns a positive correlation of $\tau = 0.543$, which is not a high correlation value, but it is statistically significant for any $\alpha$ considered with a p-value of $0.000$.

The Wilcoxon test for means does reject the null hypothesis $E(X) = E(\widehat{X})$, as the p-value obtained is lower than $\alpha = 0.05$. In fact, this test rejects the equality of the means followed by the two data sets with a strong level of significance $p - value = 0.000$. Therefore, it is concluded that $E(X)$ is significantly different from $E(\widehat{X})$.

All measures of distance suggest that *Lisboa* estimates to be worst than the *Centro* estimates, as *Lisboa* returned higher values with the measures $d_{r=2}$, $d_{r=1}$, $d_{STS}$ and $DTW$.

The $P_{NISE}$ value is zero as the $n_{lag} = 0$ as well. The $M_{NISE}$ indicates that the two data series are deviated from one another in $25.6\%$, which indicates the bad adjustment of the estimated data set.

The shape component $S_{NISE}$ for this case was $0.301$. For this case, the $PDWP$ returned a not so severe percentage distortion of the warping path, with $PDWP = 0.158$.

**Error Based Measures**

| | |
|---|---|
| $ME$ | 1.021 |
| $MAE$ | 2.824 |
| $MSE$ | 11.910 |
| $RMSE$ | 3.451 |
| $U_1$ | 0.101 |
| $U_2$ | 0.193 |
| $MPE$ | 0.005 |
| $MAPE$ | 0.157 |
| $MASE$ | 3.077 |

**Information Theory Measures**

| | | |
|---|---|---|
| $H_S$ | $H_S(X)$ | 3.148 |
| | $H_S(H, \widehat{X})$ | 3.860 |
| $I(X, \widehat{X})$ | | 2.476 |
| $NMI(X, \widehat{X})$ | | 0.786 |
| $Id(X, \widehat{X})$ | | 0.213 |

**Distance Based measures**

| | |
|---|---|
| Euclidean distance | 17.256 |
| Manhattan distance | 70.590 |
| $d_{STS}$ | 6.781 |
| DTW | 7.226 |

**Distortion Path Measures**

| | |
|---|---|
| DWP | 3.316 |
| PDWP | 0.219 |

**Parametric Tests**

| | | |
|---|---|---|
| Pearson | $r$ | 0.896 |
| | p-value | 0.000 |
| $R^2$ | | 0.802 |
| $\rho(n)$ | $\rho(n_{lag})$ | 0.896 |
| | $n_{lag}$ | 0 |

**Nonparametric Tests**

| | | |
|---|---|---|
| Sign test | TS | 13 |
| | $CR_{\alpha=0.05}$ | $[0, 7.6[\cup]17.4, 25]$ |
| Spearman | $r_S$ | 0.898 |
| | $CR_{\alpha=0.05}$ | $]0.337, +\infty[$ |
| Kendall | $\tau$ | 0.783 |
| | p-value | 0.000 |
| Wilcoxon | TS | 202 |
| | p-value | 0.299 |

**Combined measures**

| | | |
|---|---|---|
| S&G | $M_{S\&G}$ | 0.105 |
| | $P_{S\&G}$ | 0.318 |
| | $C_{S\&G}$ | 0.335 |
| Russel | $M_R$ | 0.079 |
| | $P_R$ | 0.318 |
| | $C_R$ | 0.326 |
| NISE | $M_{NISE}$ | $-0.084$ |
| | $P_{NISE}$ | 0.000 |
| | $S_{NISE}$ | 0.104 |
| | $C_{NISE}$ | 0.020 |

Table 5.4: Performance criteria summary for the number of firms in *Alentejo*.

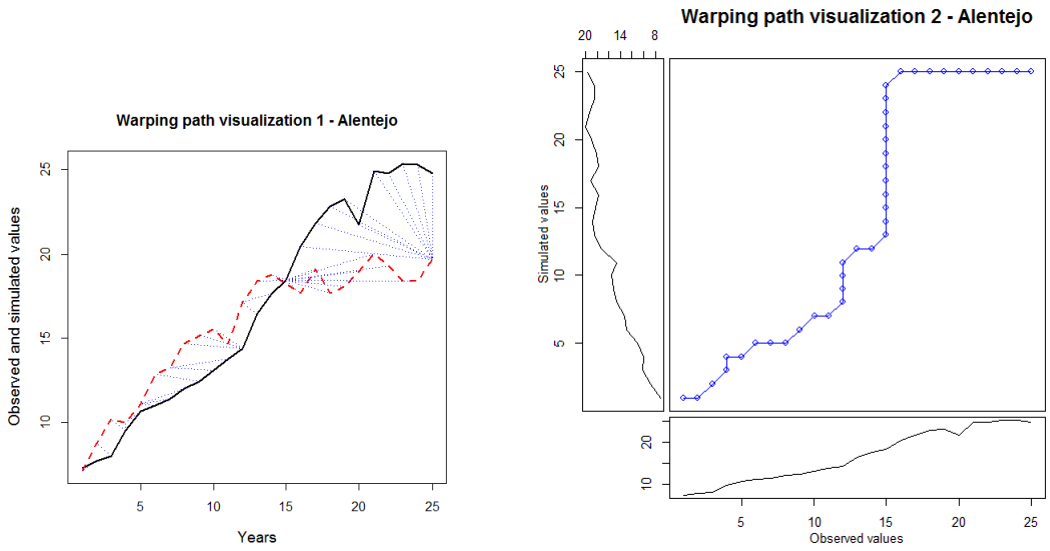


Figure 5.6: Warping path visualization for the number of firms in *Alentejo*.

### 5.2.4 Number of firms in *Alentejo*

The performance criteria obtained for the experiment conducted to simulate the number of firms in *Alentejo* are presented in Table 5.2.4. The Warping path for this experiment is shown in Figure 5.2.4.

The measures of $ME$, $MPE$, $M_{S\&G}$ and $M_R$ have positive values, indicating that in average the estimates for *Alentejo* return values below the respective observations.

The measure $U_1$ indicates that this forecast is moderately accurate, with a value of $0,101$. A similar conclusion is made with the measure $U_2$, which returned a value of $0.193$. These results suggest, however, further calibration of the model.

The estimated data set for *Alentejo* returned the lowest values of $U_1$, $U_2$, MAPE and MPE over the five features analyzed in this section. Considering these criteria, the forecast of the number of firms for *Alentejo* is the best over the five features estimated.

The value obtained with $MASE$ indicates a different conclusion, as the *Alentejo* estimates returned a $MASE$ value of $3.077$, and the best $MASE$ obtained over the five features in study is relative to *Algarve* with $2.561$. The measure $MASE$ tends to return higher values for estimations with higher absolute error values, unlike $MPE$ and $MAPE$ that are based on the percentage error. $MASE$ is based on the absolute value of the errors, and it returns a mean absolute error that is not dependent on scale. As it is possible to see, the $MAE$ for *Alentejo* e higher than the $MAE$ for *Algarve* (although they are not comparable), and this is the reason why $MASE$ provides a distinct conclusion over $MAPE$. It is not possible to evaluate whether $MASE$ is a better performance index over $MAPE$, as they are based on distinct definitions of error.

The conclusions concerning the information theory measures for this experiment are similar to the ones outlined in the former experiments.

The results obtained with Pearson's correlation test indicate that the estimated model is strongly correlated with the real observations, with $r = 0.896$ and p-value of $0.000$. The $R^2$ of $0.802$ indicates that the variability explained by the model for the *Alentejo* case is elevated. The cross correlation do not identify any temporal lag on the estimated data set.

The Sign test does not reject the null hypothesis of equality of medians among the two data series tested, meaning that for a statistical significance of $\alpha = 0.05$, the medians of $X$ and $\widehat{X}$ may be considered similar.

The Spearman's rank correlation indicates a strong and positive correlation, with $r_S = 0.898$. This result is statistically significant considering both $\alpha = 0.05$ or $\alpha = 0.01$ ($CR_{\alpha=0.01} = ]0.466, +\infty[$). As $r_S \in CR_{\alpha=0.01}$, the null hypothesis of $r_S = 0$ is rejected.

The Kendall's $\tau$ obtained for this paired data set indicates a high and positive correlation with $\tau = 0.783$. This correlation is statistically significant for any $\alpha$ considered, with a p-value of $0.000$.

The Wilcoxon Signed Rank Test for means does not reject the null hypothesis $E(X) = E(\widehat{X})$, with a p-value of $0.299$, that is considerably higher than $\alpha = 0.05$. Therefore, the null hypothesis

**Error Based Measures**

| | |
|---|---|
| $ME$ | $-0.853$ |
| $MAE$ | 2.119 |
| $MSE$ | 8.765 |
| $RMSE$ | 2.961 |
| $U_1$ | 0.109 |
| $U_2$ | 0.227 |
| $MPE$ | $-0.079$ |
| $MAPE$ | 0.232 |
| $MASE$ | 2.561 |

**Information Theory Measures**

| | | |
|---|---|---|
| $H_S$ | $H_S(X)$ | 3.074 |
| | $H_S(X, \widehat{X})$ | 3.757 |
| $I(X, \widehat{X})$ | | 2.373 |
| $NMI(X, \widehat{X})$ | | 0.772 |
| $Id(X, \widehat{X})$ | | 0.228 |

**Distance Based measures**

| | |
|---|---|
| Euclidean distance | 14.803 |
| Manhattan distance | 52.994 |
| $\mathrm{d}_{STS}$ | 15.626 |
| DTW | 6.451 |

**Distortion Path Measures**

| | |
|---|---|
| DWP | 1.316 |
| PDWP | 0.087 |

**Parametric Tests**

| | | |
|---|---|---|
| Pearson | $r$ | 0.904 |
| | p-value | 0.000 |
| $R^2$ | | 0.818 |
| $\rho(n)$ | $\rho(n_{lag})$ | 0.904 |
| | $n_{lag}$ | 0 |

**Nonparametric Tests**

| | | |
|---|---|---|
| Sign test | TS | 16 |
| | $CR_{\alpha=0.05}$ | $[0, 7.6[\cup]17.4, 25]$ |
| Spearman | $r_S$ | 0.915 |
| | $CR_{\alpha=0.05}$ | $]0.337, +\infty[$ |
| Kendall | $\tau$ | 0.774 |
| | p-value | 0.000 |
| Wilcoxon | TS | 114 |
| | p-value | 0.200 |

**Combined measures**

| | | |
|---|---|---|
| S&G | $M_{S\&G}$ | $-0.071$ |
| | $P_{S\&G}$ | 0.318 |
| | $C_{S\&G}$ | 0.326 |
| Russel | $M_R$ | $-0.060$ |
| | $P_R$ | 0.318 |
| | $C_R$ | 0.323 |
| NISE | $M_{NISE}$ | $-0.072$ |
| | $P_{NISE}$ | 0.000 |
| | $S_{NISE}$ | 0.096 |
| | $C_{NISE}$ | 0.024 |

Table 5.5: Performance Criteria summary for the number of firms in *Algarve*.

(equality of the means followed by the two data series) is kept.

The distance measures are considerably lower for the *Alentejo* experiment than the observed values with *Norte*, *Centro* and *Lisboa* experiments.

The $P_{NISE}$ value is zero as the $n_{lag}$ is zero as well. The magnitude deviation indicated by $M_{NISE}$ is only $8.4\%$, which suggests that the Alentejo estimation is close to the respective historic record. In fact, Alentejo had the second best result of $M_{NISE}$, being Algarve the best over the five experiments.

The assessment of shape dissimilarity with $S_{NISE}$ returned a moderate result of $10.4\%$. This result suggests that the patterns are reasonably similar. The pattern similarity assessment with $PDWP$ returned the value $0.219$, indicating a worst assessment of shape similarity over the two patterns than $S_{NISE}$.

### 5.2.5 Number of firms in *Algarve*

The performance indexes obtained for the number of firms in *Algarve* are presented in Table 5.2.5. The Warping path for this experiment is shown in Figure 5.2.5.

The estimated number of firms for *Algarve* returned negative values of $ME$, $MPE$, $M_{S\&G}$

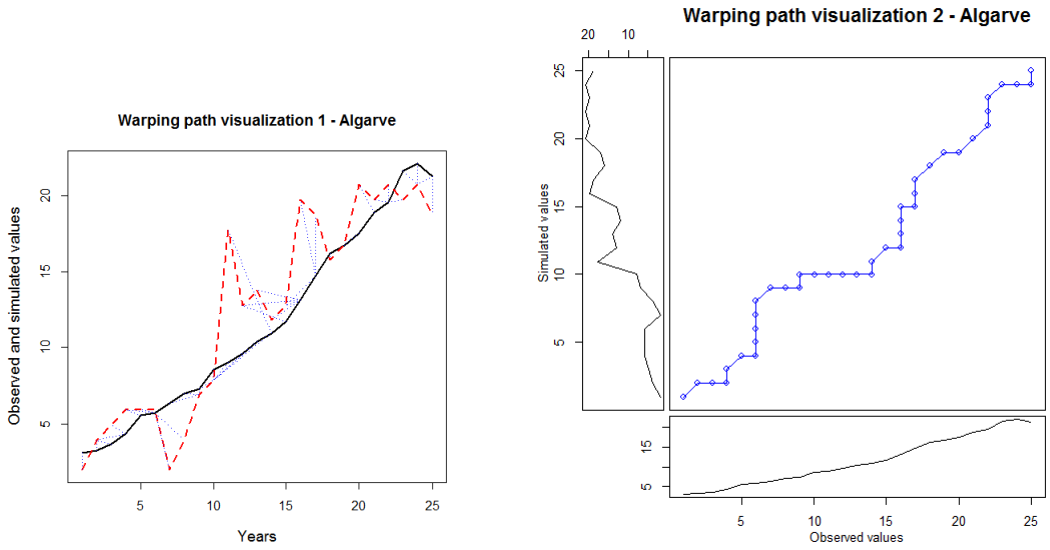


Figure 5.7: Warping path visualization for the number of firms in *Algarve*.

and $M_R$, meaning that the model tends to estimate higher values of this feature than the correspondent historic.

The measures $U_1$ and $U_2$ indicate a moderate accuracy of the results obtained for *Algarve*.

Besides the low negative value of $MPE$, the $MAPE$ value indicates that the estimations are in average deviated from the corresponding observations in $23.2\%$ (note that $MPE$ offsetts positive and negative errors).

*Algarve* has the lowest values of MASE and MAE over the five features studied. MASE is based on the absolute errors incurred by the estimation, while MAPE is based on the percentage errors obtained. For this reason, the performance criteria MASE and MAPE returned different conclusions when comparing *Alentejo* and *Algarve* experiments.

Once again, the information theory measures suggest similar conclusions as the remainder experiments. In fact, for the *Algarve* experiment, information theory measures returned the severest accuracy assessments.

The parametric tests and measures returned best results for *Algarve* over the remainder experiments analyzed. The Pearson's correlation coefficient for this case was $r = 0.904$ with a p-value of $0.000$, indicating a strong and significant positive correlation. The variability explained by the model is $0.818$ out of $1$, and no temporal lag was identified with the cross correlation test.

The $TS$ obtained with the Sign test is not within the critical region, considering a statistical significance of $\alpha = 0.05$. Therefore, the null hypothesis testing the equality of $X$ and $\widehat{X}$ medians is not rejected, and these medians are considered similar.

The Spearman's rank of correlation indicates that the two data sets are highly and positively correlated, considering both $\alpha = 0.05$ or $\alpha = 0.01$ ($CR_{\alpha=0.01} = ]0.466, +\infty[$), as $r_S \in CR$ for

both cases, the null hypothesis $r_S = 0$ is always rejected.

The Kendall's $\tau$ obtained suggests that the two data sets are positively correlated, with $\tau = 0.774$. This correlation is statistically significant for any $\alpha$ considered.

The Wilcoxon test for means returned a p-value of $0.200$, that is higher than $\alpha = 0.05$ and consequently the null hypothesis of $E(X) = E(\widehat{X})$ is not rejected. Thus the assumption of equality of the means followed by the two data sets under analysis is kept.

Concerning the distance measures, the *Algarve* returned the lowest values over the five experiments. This means that considering the distance measures, *Algarve* would be the best forecast among the five analysed.

For the *Algarve* experiment, the $P_{NISE}$ value is zero as the $n_{lag}$ is zero as well. The assessment of magnitude deviation with $M_{NISE}$ returned the best result over the five cases, with $M_{NISE} = 7.2\%$. This result is in accordance with the MASE result.

The measures of shape component $S_{NISE}$ and $PDWP$ returned for *Algarve* the best results over all experiments, suggesting that this was the best estimation provided by the model concerning the pattern adjustment.

# Chapter 6

# Conclusions

A computational simulation model is designed to validate or reject the hypothesis formulated in the conceptual model. A simulation model can be accepted as valid for specific conditions only, which means that under the assumptions specified, the model reproduces well the task for which it was designed. To prove that a model reproduces well its task, the standard procedures to confirm scientific theories should be followed [41]. These procedures include the use of data from the reality under study to confront the results obtained with the model. This confront happens within the phases of *calibration* and *validation*. The outcomes from validation determine weather the hypothesis formulated in the conceptual model should be accepted or rejected.

Each phase integrating the process described of constructing a simulation model, encompasses many complex aspects. This work tried to investigate one of these aspects: the adequate performance criteria to use under a validation process. The performance criteria considered in this work are those adequate to assess the goodness of fit between paired data samples, always considering the paired data sample to be the observed and simulated data sets.

The main contributions of this thesis are (i) the bibliographic overview on performance criteria, (ii) the proposal of two new performance criteria, the Warping Path Distortion $WPD$ and the Percentage Warping Path Distortion $PWPD$, and (iii) the comparative analysis of the criteria reviewed under practical experiments.

Next the main advantages and drawbacks identified in this work for the criteria reviewed are detailed.

The measures Mean Error $ME$, Mean Absolute Error $MAE$, Mean Square Error $MSE$ and Root Mean Square Error $RMSE$ are useful when applied to assess different runs on the same feature of the same case study. Only in that situation can these measures be compared.

The Theil's coefficient for forecast accuracy $U_1$ returns values included within the range $[0, 1]$ and indicates good estimates for values close to zero. This is a scale independent measure that can be directly compared over different case studies. However, $U_1$ does not have an accuracy frontier to precisely define weather a validation result should be rejected or accepted. Therefore,

the analysis made with this measure are always subjective when assessing a single estimation. $U_1$ is more advantageous when applied to compare the accuracy over different estimates.

The coefficient for forecast quality $U_2$ returns values within the range $[0, +\infty[$, being the value $0$ indicative of a perfect forecast. Values lower than $1$ indicates that the estimation obtained has lower standard error as the naive no-change extrapolation. If a value larger or equal to $1$ is obtained for the $U_2$ criterion, the model should be rejected as in that case the standard error obtained with the forecast would be worst than the simples no-change extrapolation.

The sign obtained with the measures $ME$, Mean Percentage Error $MPE$, Sprague and Gear magnitude error $M_{S\&G}$ and Russel's magnitude error $M_R$ is negative when the estimated data returns in average higher values than the ones observed, and positive otherwise. The identification of magnitude divergences is a good starting point to improve future calibration on models. The calibration of the model concerning the magnitude should be made together with the visualization of the corresponding plots. The visual observation is important as, for example, in the *Norte* experiment, negative values were obtained for these measures, and the calibration needed in this case is more related with an improvement on the resultant slant than the blind decrease of the estimated values.

The criterion Mean Absolute Scaled Error $MASE$ returns values within the range $[0, +\infty[$. Since this measure is not normalized, its applicability is more useful to compare results over different estimations than to assess the goodness of fitness of an isolated estimation.

$MASE$ is based on the absolute errors incurred by the estimation, while Mean Absolute Percentage Error $MAPE$ is based on the percentage errors obtained. For this reason, the comparison of the quality of adjustment between *Alentejo* and *Algarve* returned different conclusions with $MASE$ and $MAPE$.

The information theory measures returned the severest validation results, indicating that all the five estimated data sets are very poor. The Normalized Mutual Information $NMI$ and Normalized information distance $Id$ showed to be valuable performance criteria due to its easiness of interpretation: both return values within the range $[0, 1]$ and can be interpreted as information percentages.

The parametric test of Pearson's correlation $r$, and the parametric measures Coefficient of determination $R^2$ and Cross correlation $\rho(n)$ should not be conducted without demonstrating the normality of the samples' population. These criteria were presented for the five experiments with the relaxation of this main assumption, in order to provide a fuller comparative criteria assessment.

As explained by [25], $r$ and $R^2$ are insensitive to additive and proportional differences between simulated and observed homologous elements. This is the main drawback of these two criteria.

The distance between the two data sets (considered as two points points defined within a $N$-dimensional space) is a valuable performance criteria. The Euclidean distance $d_{r=2}$ and the Manhattan distance $d_{r=1}$ indicate better estimations for values close to zero. When a single experiment is provided, the distance value does not provide much information, similarly to the reasoning

explained for $ME$, $MAE$, $MSE$ and $RMSE$.

The Short Time Series distance $d_{STS}$ returns a measurement on the average amplitude divergence over consecutive temporal moments between two data sets. To be precise, it returns the square root of the sum of the squared differences of the slopes between temporal moments. The complex interpretation of this measure is found to be its main disadvantage. Moreover, $d_{STS}$ is insensitive to additive differences between the two data sets under assessment. For example, considering the situation where $\widehat{x}_i = x_i + 1, \forall i$, the value $d_{STS} = 0$ would be obtained, which is the reference value for a perfect fit.

Another distance measure revised is the Dynamic Time Warping $DTW$. The comparison between the measures $DTW$ and $d_{r=2}$ is possible. This comparison can only be made when: (i) the two data sets $X$ and $\widehat{X}$ have the same length, (ii) $DTW$ is calculated as defined in equation (3.50), and (iii) the cost function used to calculate matrix $\mathbf{A}$ (3.48) is the squared error $d(x_i, \widehat{x}_j) = (x_i - \widehat{x}_j)^2$. When this three conditions are ensured, $DTW$ values closer to $d_{r=2}$ indicate a higher similarity between the two patterns. However, this comparison would not indicate how much similar the two patterns are, due to the scale-dependency of both measures.

The $DTW$ values analyzed in chapter 5 used the euclidean distance $d(x_i, \widehat{x}_j) = d_{r=2}(x_i, \widehat{x}_j)$ as the cost function to calculate the elements of matrix $\mathbf{A}$ (3.48), from where the values of $DTW$ reported were not comparable with the $d_{r=2}(X, \widehat{X})$.

The main drawback identified for the $DTW$ is that it is dependent on the scale and on the length of the observations analyzed. This constrains the assessment of forecasts relating to different case studies with $DTW$.

The Warping Path Distortion $WPD$ is based on the warping path constructed within the $DTW$ algorithm. $WPD$ returns the average distance between the Warping Path and the corresponding diagonal positions. The measure $WPD$ has the advantage of being scale independent, from where it may be used to compare the accuracy of different forecasts. $WPD$ main disadvantage is that it depends on the number of observations within the data sets $N$. This means that it should only be used to compare different case studies provided the length of the data sets used is similar. Values of $WPD$ close to zero indicate a better fit. The maximum value observable with $WPD$ vary depending on $N$ and on the number of elements integrating the respective Warping Path $K$.

The disadvantage identified with $WPD$ is surpassed by the Percentage Warping Path Distortion $PWPD$. $PWPD$ is not dependent on the observations scale nor the number of observations. The values returned by $PWPD$ are easily interpreted as percentages, as they vary within the range $[0, 1]$. The perfect pattern similarity returns a $PWPD$ value of zero. The worst pattern dissimilarity would return the asymptotic $PWPD$ value of one, as it was illustrated in Figure 4.1. Although the two measures proposed behaved as expected under the experiments analyzed, further applicability of these measures to benchmark data sets is necessary to provide a proper conclusion on its quality.

The Sprague and Gear phase error $P_{S\&G}$ and Russel's phase error $P_R$ are suitable to assess lag on series with sinusoidal behavior, or wave form, as explained in [45]. Therefore, the application of these measures to the case study of Portuguese firms are meaningless.

The Normalized Integral Square phase error $P_{NISE}$ is related with the results obtained with $\rho(n)$. When the number of temporal periods of lag $n_{lag}$ obtained with $\rho(n)$ is different from 0, $P_{NISE}$ returns values different from zero. When $n_{lag} = 0$, the $P_{NISE} = 0$ as well.

The Sprague and Gear magnitude error $M_{S\&G}$ and Russel's magnitude error $M_R$ return values within the range $]-1, 1[$. They are interpreted as the percentage magnitude discrepancy between the series under analysis, with sensitivity to the sign of the discrepancy (similar reasoning as $MPE$).

Concerning the Normalized Integral Square magnitude error $M_{NISE}$, it returns values within the range $[-1, 0]$, and is interpreted as an absolute magnitude percentage deviation (besides the negative sign). The values obtained with $M_{NISE}$ have similar interpretation to the $MAPE$ measure.

The shape component Normalized Integral Square shape error $S_{NISE}$ returns values within the range $[0, 1]$, being 0 the best possible result indicating a good pattern similarity and 1 the worst possible result indicating a bad similarity. The measure $PDWP$ returns values within the same range and with the same interpretation as $S_{NISE}$.

It would be almost impossible to encompass all performance criteria referenced in literature. Nevertheless this thesis is at least a starting point to guide the choice of an adequate performance criterion for a specific validation model.

Future research may include the study on how to apply the Simple String Distance Metric, suggested in [12], to a general process of validation. This measure was suggested under the context of validation of DNA simulated patterns. Another interesting issue to explore under this context, is the use of combined graphical and statistical approaches, that is summarily described in the review paper on validation [6]. Another relevant topic for future research would be the comparative assessment on the computational effort of performance criteria.

# Bibliography


[1] C. Adami. The use of information theory in evolutionary biology. *Annals of the New York Academy of Sciences*, 1256:49–65, 2012.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.

[3] Hirotugu Akaike. Likelihood of a model and information criteria. *Journal of Econometrics*, 16:3–14, 1981.

[4] J.S. Armstrong and F. Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8:69–80, 1992.

[5] Vic Barnett. *Sample Survey Principles and Methods*. John Wiley & Sons, 2003.

[6] Gianni Bellocchi, Mike Rivington, Marcello Donatelli, and Keith Matthews. Validation of biophysical models: issues and methodologies. a review. *Agronomy for Sustainable Development*, 30(1):109–130, 2009.

[7] R. J. Bessa. Treino on line de redes neuronais com critérios de informação aplicado à previsão eólica. Master's thesis, Faculdade de Economia da Universidade do Porto, 2008.

[8] Friedhelm Bliemel. Theil's forecast accuracy coefficient: A clarification. *Journal of Marketing Research*, 10:444–446, 1973.

[9] Franky Kin Pong Chan and Ada Wai chee Fu. Haar wavelets for efficient similarity search of time series: With and without time warping. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):686–705, 2003.

[10] Selina Chu, Eamonn Keogh, David Hart, and Michael Pazzani. Iterative deepening dynamic time warping for time series. In *Proceedings of the 2nd SIAM international conference on data mining.*, 2002.

[11] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley & sons, 2nd edition, 1980.

[12] J.E. Cook and A.L. Wolf. Toward metrics for process validation. In *Proceedings of the Third International Conference on the Software Process*, pages 33–44, 1994.

[13] Kevin K. Dobbin and Richard M. Simon. *Optimally splitting cases for training and testing high dimensional classifiers*, volume 4. 2011.

[14] Giorgio Fagiolo, Alessio Moneta, and Paul Windrum. A critical guide to empirical validation of agent based models in economics: Methodologies, procedures, and open problems. *Computational Economics*, 30:195–226, 2007.

[15] M. Gardner. Mathematical games the fantastic combinations of John Conway's new solitaire game of life. *Scientific American*, 223, 1970.

[16] Nigel Gilbert and Klaus G. Troitzsch. *Simulation for the Social Scientist*. Open University Press, Buckingham, Philadelphia, 1999.

[17] D. Gujarati. *Econometria*. McGrow Hill, 2nd edition, 1996.

[18] Rob J. Hyndman. Another look at forecast accuracy metrics for intermittent demand. *Foresight*, 4:43–46, 2006.

[19] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22:679–688, 2006.

[20] X. Jiang and S. Mahadevan. Bayesian cross entropy methodology for optimal design of validation experiments. *Measurement Science and Technology*, 17(7):1895–1908, 2006.

[21] V. Klemeš. Operational testing of hydrological simulation models. *Hydrological Sciences*, 31(1):13–24, 1986.

[22] L.F. Konikow and J.D. Bredehoeft. Ground water models cannot be validated. *Validation of Geo-hydrological Models*, 15(1):75–83, 2003.

[23] Jouni Kuha. AIC and BIC comparisons of assumptions and performance. *Sociological Methods & Research*, 33(2):188–229, 2004.

[24] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962.

[25] David R. Legates, McCabe Jr., and Gregory J. Evaluating the use of goodness of fit measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1):233–241, 1999.

[26] Beitao Li, Edward Chang, and Ching-Tung Wu. DPF a perceptual distance function for image retrieval. In *IEEE 2002 International Conference on Image Processing*, 2002.

[27] Beitao Li, Edward Chang, and Yi Wu. Discovery of a perceptual distance function for measuring image similarity. *Multimedia Systems*, 8, 2003.

[28] Wentian Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60(5–6):823–837, 1990.

[29] T. Warren Liao. Clustering of time series data a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[30] Iain L. MacDonald and Walter Zucchini. *Hidden Markov and Other Models for Discrete-valued Time Series*. Springer, 1997.

[31] M.D. McDonnell, S. Ikeda, and J.H. Manton. An introductory review of information theory in the context of computational neuroscience. *Biological Cybernetics*, 105(1):55–70, 2011.

[32] Tom M. Mitchell. *Machine Learning*. The McGraw Hill Companies, Inc.

[33] Bruce Mizrach. The distribution of the Theil U statistic in bivariate normal populations. *Economics Letters*, 38(2):163–167, 1992.

[34] C.S. Moller Levet, F. Klawonn, K.H. Cho, and O. Wolkenhauer. Fuzzy clustering of short time series and unevenly distributed sampling points. In *Proceedings of the 5th International Symposium on Intelligent Data Analysis, Berlin, Germany.*, 2003.

[35] N. Oreskes, K. Shrader-Frechette, and K. Belitz. Verification, validation, and confirmation of numerical models in the earth sciences. *Science, New Series*, 263(5147):641–646, 1994.

[36] Sanghyun Park and Vijay S. Pande. Validation of Markov state models using shannon's entropy. *The Journal of Chemical Physics*, 124(054118):1–5, 2006.

[37] K. Popper. *The Logic of Scientific Discovery*. London: Hutchingson Co, 1959.

[38] D. Posada and T.R. Buckley. Model selection and model averaging in phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808, 2004.

[39] J. C. Principe. *Information Theoretic Learning Renyi's Entropy and Kernel Perspectives*. Springer, 2010.

[40] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[41] J.C. Refsgaard and H.J. Henriksen. Modelling guidelines terminology and guiding principles. *Advances in Water Resources*, 27:71–82, 2004.

[42] R.G. Sargent. Validation and verification of simulation models. In *Proceedings of the 2004 Winter Simulation Conference*, 2004.

[43] H. Sarin, M. Kokkolaras, G. Hulbert, P. Papalambros, S. Barbat, and R. J. Yang. A comprehensive metric for comparing time histories in validation of simulation models with emphasis on vehicle safety applications. In *Proceedings of DETC 08, ASME 2008 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2008.

[44] H. Sarin, M. Kokkolaras, G. Hulbert, P. Papalambros, S. Barbat, and R. J. Yang. Comparing time histories for validation of simulation models: Error measures and metrics. *Journal of Dynamic Systems, Measurement, and Control*, 132(6):1–10, 2010.

[45] Leonard E. Schwer. Validation metrics for response histories: perspectives and case studies. *Engineering with Computers*, 23(4):295–309, 2007.

[46] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, 1948.

[47] P. Sprent. *Applied Nonparametric Statistical Methods*. Chapman & Hall, 2nd edition, 1993.

[48] Leon S. Sterling and Kuldar Taveter. *The Art of Agent Oriented Modeling*. The MIT Press.

[49] Richard Taylor. Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, 6(1):35–39, 1990.

[50] Henry Theil. *Economic Forecasts and Policy*. North Holland Pub. Co., 1st edition, 1965.

[51] Henry Theil. *Applied Economic Forecasts*. North Holland Pub. Co., 1st edition, 1966.

[52] C. Turkay, E. Koc, and S. Balcisoy. Integrating information theory in agent based crowd simulation behavior models. *The Computer Journal*, 54(11):1810–1820, 2011.

[53] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

[54] David L. Weakliem. A critique of the bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3):359–397, 1999.

[55] Cort J. Willmott. On the validation of models. *Physical Geography*, 2(2):184–194, 1981.

[56] Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30:79–82, 2005.

[57] J. H. Zar. Significance testing of the Spearman rank correlation. *Journal of the American Statistical Association*, 67, 1972.